the CATO JOURNAL

AN INTERDISCIPLINARY JOURNAL OF PUBLIC POLICY ANALYSIS

ARTICLES

Charlotte Twight — Evolution of Federal Income Tax Withholding: The Machinery of Institutional Change

Stephen C. Erickson — The Entrenching of Incumbency: Reelections in the U.S. House of Representatives, 1790–1994

John C. Moorhouse — Competitive Markets for Electricity Generation

John J. Siegfried — Sports Player Drafts and Reserve Systems

Michael D. Bordo and Anna J. Schwartz — The Performance and Stability of Banking Systems under "Self-Regulation": Theory and Evidence

George Selgin — Bank Self-Regulation: Comment on Bordo and Schwartz

Ralph Chami and Jeffrey H. Fischer — Community Banking, Monitoring, and the Clinton Plan

COMMUNICATIONS

Harold J. Brumm and Dale O. Cloninger — The Drug War and the Homicide Rate: A Direct Correlation?

Herschel Grossman — Scholarships: Need or Merit?

VOLUME 14 NUMBER 3 • PUBLISHED BY THE CATO INSTITUTE

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the CATO JOURNAL

VOLUME 14 NUMBER 3 • PUBLISHED BY THE CATO INSTITUTE • WINTER 1995

Editor
JAMES A. DORN

Managing Editor
MONIKA VANBERG

Book Review Editor
SHELDON RICHMAN

EDITORIAL BOARD

TERRY L. ANDERSON
Montana State University

PETER H. ARANSON
Emory University

CHARLES W. BAIRD
California State University at Hayward

T. E. BORCHERDING
Claremont Graduate School

YALE BROZEN
University of Chicago

JAMES M. BUCHANAN
Center for Study of Public Choice
George Mason University

RICHARD A. EPSTEIN
University of Chicago

MARILYN R. FLOWERS
Ball State University

JAMES GWARTNEY
Florida State University

STEVE H. HANKE
Johns Hopkins University

ROBERT HIGGS
The Independent Institute

JERRY L. JORDAN
Federal Reserve Bank of Cleveland

LEONARD P. LIGGIO
Atlas Foundation

ANTONIO MARTINO
LUISS University

DONALD N. McCLOSKEY
University of Iowa

RICHARD B. McKENZIE
University of California, Irvine

WILLIAM A. NISKANEN
Cato Institute

SAM PELTZMAN
University of Chicago

GEORGE A. SELGIN
University of Georgia

JUDY SHELTON
Marshall, Virginia

VERNON L. SMITH
University of Arizona

ROLAND VAUBEL
Mannheim University

LAWRENCE H. WHITE
University of Georgia

LELAND B. YEAGER
Auburn University

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The CATO JOURNAL (ISSN 0273-3072) is published in the spring/summer, fall, and winter by the Cato Institute, 1000 Massachusetts Ave., NW, Washington, D.C. 20001-5403. Annual subscriptions are $24.00 for individuals and $50.00 for institutions. Single copies are $8.00. Two-year and three-year subscriptions are $48.00 and $72.00, respectively, for individuals, and $100.00 and $150.00, respectively, for institutions. Foreign subscribers should add $5.00 per year for regular delivery and $10.00 per year for airmail delivery.

Correspondence regarding subscriptions, changes of address, procuring back issues, advertising and marketing matters, etc., should be addressed to the Publications Department. All other correspondence, including requests to quote or reproduce material, should be addressed to the editor.

Unsolicited manuscripts cannot be returned and will be acknowledged only if accompanied by a stamped, self-addressed envelope. There is no submission fee. Articles should focus on specific public policy issues. Accepted manuscripts must conform to the CATO JOURNAL's style requirements. A style sheet is available on request.

Authors should submit three copies of their papers or communications to the editor. An abstract of no more than 250 words should also be included to aid referees. Manuscripts should have wide margins and be double-spaced throughout, including footnotes.

Unsolicited book reviews cannot be returned and will be acknowledged only if they are accepted for publication. They should conform to the above requirements and should provide an in-depth analysis of major studies of public policy. Reviews should range from 800 to 1,000 words.

The views expressed by the authors of the articles are their own and are not attributable to the editor, the editorial board, or the Cato Institute.

The CATO JOURNAL is indexed and abstracted in the *Journal of Economic Literature, Current Contents/Social and Behavioral Sciences, Social Sciences Citation Index, America: History and Life, Historical Abstracts,* and the **ABI/INFORM** data base. The CATO JOURNAL is available on microfilm from University Microfilms International, Ann Arbor, MI 48106.

Printed in the United States of America.

Copyright © 1995 by the Cato Institute.
All rights reserved.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

CATO JOURNAL • VOLUME 14 NUMBER 3 • WINTER 1995

ARTICLES

Charlotte Twight
Evolution of Federal Income Tax Withholding: The Machinery of Institutional Change 359

Stephen C. Erickson
The Entrenching of Incumbency: Reelections in the U.S. House of Representatives, 1790–1994 397

John C. Moorhouse
Competitive Markets for Electricity Generation 421

John J. Siegfried
Sports Player Drafts and Reserve Systems 443

Michael D. Bordo and Anna J. Schwartz
The Performance and Stability of Banking Systems under "Self-Regulation": Theory and Evidence 453

George Selgin
Bank Self-Regulation: Comment on Bordo and Schwartz 481

Ralph Chami and Jeffrey H. Fischer
Community Banking, Monitoring, and the Clinton Plan 493

COMMUNICATIONS

Harold J. Brumm and Dale O. Cloninger
The Drug War and the Homicide Rate: A Direct Correlation? 509

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Herschel Grossman
Scholarships: Need or Merit? 519

BOOK REVIEWS

Out of Work: Unemployment and Government in Twentieth-Century America
Richard K. Vedder and Lowell E. Gallaway

Reviewed by James T. Bennett 527

Federalism and Free Trade
Jean-Luc Migué

Reviewed by Dwight R. Lee 529

Why Perestroika Failed: The Politics and Economics of Socialist Transformation
Peter J. Boettke

Reviewed by Scott Shane 531

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Evolution of Federal Income Tax Withholding: The Machinery of Institutional Change

Charlotte Twight

Taxes are the backbone of any politico-economic regime. Constraints on a government's power to tax are constraints on its power to act. Focusing on the legalization of mandatory federal income tax withholding through the Current Tax Payment Act of 1943, this article examines forces that have eroded constraints on the U.S. government's power to tax.

The central questions this article seeks to answer are how, why, and to what effect—despite preponderant public opposition to universal income tax withholding between 1914 and 1942—mandatory withholding was established in 1943, and sustained thereafter. It is an important question, for withholding is the paramount administrative mechanism enabling the federal government to collect, without significant protest, sufficient private resources to fund a vastly expanded welfare state. U.S. government officials themselves now view withholding as "the cornerstone of the administration of our individual income tax" (U.S. House Ways & Means Comm. Hearings 1982: 162, 165). This article explores (1) historical conditions that led people to accept withholding of federal taxes on wage and salary income; (2) the politico-economic function of income tax withholding; and (3) the consistency of the U.S. income tax withholding experience with a more general economic model of institutional and ideological change.

A systematic transaction-cost framework for understanding the evolution of withholding will be suggested as a way of integrating this critical episode with other U.S. policy experiences. Historical circumstances facilitating adoption and expansion of income tax withholding

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

The author is Professor of Economics at Boise State University. She gratefully acknowledges a grant by the Earhart Foundation that facilitated this research. She wishes to thank Robert Higgs and an anonymous referee for helpful comments on an earlier draft.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

will be seen to reflect broader incentives of government officials to alter political transaction costs facing the public. Equally important, the willful alteration of political transaction costs will be viewed as supporting institutional and ideological changes that, over time, expand the publicly accepted scope of government authority.

Government manipulation of political transaction costs contributed significantly to the institutional changes and subsequent ideological transformation supporting income tax withholding. Though in 1943 the withholding mechanism was sold politically as a benefit to taxpayers, government officeholders even then widely regarded it as a means of extracting greater tax revenue. Senators and representatives spoke candidly in congressional hearings (U.S. Senate Hearings 1943: 43) of the revenues that needed "to be fried out of the taxpayers."

This article is organized as follows. I first develop a theoretical framework for analysis of federal income tax withholding. After discussing the early evolution of income tax withholding in the United States, I then consider how changing institutional contexts have influenced attempts to expand withholding. The final section of the article ponders the reversibility of current withholding mechanisms in light of the paper's theory and evidence.

Institutional and Ideological Change: A Theoretical Framework

Many scholars have identified politico-economic patterns associated with taxation. For instance, in his study of taxation in the New Deal, Mark Leff (1984) showed that the rhetoric surrounding tax policy often serves a symbolic function inconsistent with actual tax policy. John Witte (1985) emphasized that across a broad span of income tax history changes in tax law have tended to proceed *incrementally* and therefore to generate *complexity*. Dall Forsythe (1977), studying U.S. tax policy during 1781–1833, suggested that tax policy is shaped by recurrent political patterns, including "normal politics," "regime politics," "environmental crises" such as war or depression, and "authority crises" such as the Civil War in which the regime's ability to govern is challenged. He viewed these patterns as helping to explain why similar policy initiatives sometimes generate quite different political outcomes, arguing (1977: 122), for instance, that

> if the elite can successfully establish its definition of a situation as a crisis, it can undertake without direct opposition activities which might otherwise be considered gross violations of regime boundaries.

All of the above-cited conceptualizations of the emergence of tax policy may be subsumed by a broader explanation grounded in the

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

phenomenon of government manipulation of political transaction costs. As employed here, the phrase "political transaction costs" denotes transaction costs involved in making political decisions on issues that influence the scope of government authority.[1] Those transaction costs include not only costs of obtaining information regarding an authority-influencing issue, but also costs of taking collective action to support one's policy preferences. A transaction-cost-manipulation model identifies the endogeneity of a broad range of transaction costs affecting the accepted scope of government authority, and it endeavors to specify both the circumstances that give rise to government augmentation of political transaction costs and the consequences of that behavior.[2] I have analyzed theoretical dimensions of this model in greater detail elsewhere (Twight 1983, 1988, 1994). In the present context, I want only to summarize salient features of the model as a backdrop for this article's historical analysis of the evolution of U.S. income tax withholding policy.

Transaction-cost-manipulation theory builds on the idea that government officials, as individuals, have strong incentives to try to alter political transaction costs facing citizens and others in government. If they can raise the transaction costs to voters of opposing a policy that the officeholders (or influential constituents) favor, the officeholders' policy preferences are more likely to prevail. Clearly, if government officials make it harder for most citizens either to perceive an unwanted policy (e.g., a tax or special-interest legislation) or to organize to resist it, public opposition is less likely to materialize. In the presence of transaction-cost augmentation, political resistance to policy initiatives is thus in part a function of government officeholders' volitional choices to raise transaction costs of particular types of collective action, not solely a function of citizens' or other officeholders' preferences.

We will see in the next section that income tax withholding is both an instrument and a result of transaction-cost augmentation. Withholding's transaction-cost-increasing features and implications

[1]Alternatively, such political transaction costs could be labeled "constitutional-level" transaction costs (see Twight 1988, 1992) to emphasize their influence upon the nature and extent of government authority over private decisionmaking tolerated by the public. In this broad sense, to conceptualize them as "constitutional-level" implies only that they influence where the line is drawn between the public sector and the private sector; it does not presuppose linkage to or embodiment in a formal constitutional document. Regarding alternative loci of a constitution, see Higgs (1988: 374–75).

[2]Transaction-cost-manipulation theory thus is concerned with what may be called "contrived" political transaction costs, volitionally created by political actors for their own benefit. The residual political transaction costs that exist when all agents are attempting to minimize transaction-cost barriers to political decisionmaking will be referred to as "natural" transaction costs.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

for the growth of government were well captured by David Brinkley (as quoted in Jones 1989: 730), who stated that, with withholding

> Congress and the president learned, to their pleasure, what automobile salesmen had learned long before: that installment buyers could be induced to pay more because they looked not at the total debt but only at the monthly payments. And in this case there was, for government, the added psychological advantage that people were paying their taxes with not much resistance because they were paying with money they had never even seen.

As George Lent (1942) described it in the *Journal of Political Economy*, "the taxpayer does not have the same consciousness of parting with his income to the government," making withholding "the most 'painless' method of meeting tax liabilities." The following section explores the extent to which government decisionmakers not only understood this result *ex ante* but also used other types of transaction-cost manipulation to achieve it when they instituted income tax withholding in 1943.

Since the history of U.S. income tax withholding documented below involves extensive misrepresentation by government officeholders, and since withholding itself alters perceptions of private tax burdens, one may legitimately ask whether a transaction-cost manipulation model is concerned exclusively with use of political deception. As the preceding paragraphs imply, the answer is no. Whether or not deception is involved in passage and implementation of a law such as the Current Tax Payment Act of 1943, the institutions and constituencies thereby fostered can significantly alter the transaction costs of political resistance to present and future government policies. In short, political transaction costs involve costs of individual and collective political action broadly defined, not just information costs. Regardless of the strategies by means of which income tax withholding became law in the United States, the institutional practice of withholding affected in predictable ways the political viability of income taxation and the government policies such taxation supports.

Consider the following conceptual experiment to assess the policy neutrality of income tax withholding. Suppose that we eliminated mandatory withholding for a year and instead required taxpayers to send in checks on April 15 for the full amount of their annual federal income taxes. The likely consequences of this institutional change are consistent with government officials' beliefs (documented below) concerning the role of withholding in increasing private individuals' costs of tax resistance. Most relevant here, with government-mandated withholding, such increased costs of tax resistance would exist even under the counterfactual assumption of perfect public understanding

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

of the impact of withholding on the real burden of taxation: that is, they need not depend on continuing deception.

Transaction-cost augmentation takes myriad forms. On the collective-action side, it includes changing the locus of decisionmaking authority so as to shift the transaction-cost burden entailed in effectuating changes in the role of government,[3] changing the cost to private citizens of achieving political agreement to revise the scope of governmental authority, changing the interaction between governmental agencies to alter the cost to individuals of revising the scope of government authority, and concentrating the benefits and dispersing the harm born of government action (Twight 1988: 150, fn. 1 and 2; 1994). On the information-cost side, transaction-cost manipulation includes such strategies as semantic efforts to alter public perceptions of the costs or benefits of government activities, forms of taxation that change people's perception of the actual tax burden imposed upon them, and overt distortion of information about the nature and consequences of government activities.

Many superficially disparate forms of political or legislative behavior fit within this model, and the model helps us to predict the context-specific likelihood of politicians' use of such strategies. Leff's (1984) evidence of symbolic rhetoric as well as Witte's (1985) evidence of the roles of incrementalism and complexity thus may be viewed in this broader context. Moreover, governmental attempts to milk bogus crisis circumstances for expanded authority, as described by Forsythe (1977), involve analogous efforts to raise transaction costs to the voting public of resisting governmentally favored policy measures. In other policy contexts, diverse political strategies—such as inserting riders in omnibus bills, diffusing tax costs of public policies, expanding sovereign immunity doctrines, establishing ballot-access laws that differentially impact third parties, and using federal tax money (such as federal funding for universities and highways) as a lever to foster unrelated federal policies—all alter the transaction costs to private citizens of participating in political processes affecting the role and scope of government.[4]

The question is, under what circumstances is it more likely that government officials will choose to employ transaction-cost augmentation? In previous research (Twight 1988) I have suggested that a

[3]For example, U.S. Supreme Court redefinition of the Constitution's interstate commerce clause in ways that previously would have required constitutional amendment.

[4]The examples in the text are intended to be suggestive rather than exhaustive. For a more complete discussion of political behaviors and policies that exemplify transaction-cost augmentation, see Twight (1994).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

government official's decision to favor a transaction-cost-increasing measure is likely to be a positive function of identifiable variables, including the issue's complexity,[5] the availability of an appealing (though possibly incorrect) rationale for the measure, executive support and party support for the measure, third-party payoffs, the measure's perceived importance to constituents, and its promise of job security and perquisites for the officeholder. The official's decision is expected to be a negative function of publicity or media attention directed at the measure's transaction-cost-increasing features. The impact of time is an empirical issue because, while it counteracts complexity, it also facilitates entrenchment of beneficiary interest groups. The officeholder's ideology also is expected to play an important role, influencing him to favor measures that raise the transaction costs of opposing measures that the officeholder favors on ideological grounds.[6]

As preliminary evidence of the relevance of a transaction-cost-augmentation model in explaining the development and extension of U.S. income taxation, consider two studies. Although the authors did not analyze their results in these terms, both Carolyn Jones (1989) and Ben Baack and Edward John Ray (1985a, 1985b) provide evidence of the importance of political transaction-cost manipulation in engineering acceptance of the income tax.

Jones documented the widespread and systematic use of propaganda by U.S. government officials during World War II to quell resistance to the transformation of the income tax from a "class tax" to a "mass tax" during those years. This propaganda ranged from pressuring radio broadcasters to air "plugs" promoting income tax payment to providing story lines to magazines. However, in Jones's view (1989: 716) the "crown jewel of tax propaganda" was a Disney

[5]The complexity and appealing rationale variables require further comment. An issue's complexity may be either unavoidable (hence a "natural" transaction cost of understanding an issue) or itself a product of transaction-cost augmentation (hence a "contrived" transaction cost). Similarly, an appealing rationale may be either false (a contrived transaction cost that in turn facilitates other forms of transaction-cost augmentation) or true. In either case, the direction of impact on an officeholder's decision regarding a transaction-cost-increasing measure is as described in the text. Issue complexity, like the existence of an appealing rationale, makes it harder for citizens to perceive transaction-cost-increasing features of policy proposals. Moreover, both of these conditions allow politicians greater room to credibly claim to have made a "mistake" if negative public reaction to the measure's transaction-cost-increasing features does materialize.

[6]These relationships have proved consistent with actual U.S. policymaking experience in such apparently diverse arenas as off-budget expenditure through the Federal Financing Bank, military base closures, asbestos regulation, and U.S. Social Security legislation (see Twight 1983, 1989, 1991, 1993).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

film entitled *The New Spirit* commissioned and promoted by the U.S. Treasury Department, in which Donald Duck was informed "that it is 'your privilege, not just your duty, but your privilege to help your government by paying your tax and paying it promptly'." More than 32 million people saw the film in the first few months of 1942, and a Gallup poll reported that "37 percent felt the film had affected their willingness to pay taxes" (Jones 1989: 717). Without doubt, such government propaganda manipulated political information in ways that raised the expected marginal cost of income tax resistance.

Lest Jones's observations appear anomalous, note that the U.S. government employed income tax propaganda well before World War II. During World War I, the secretary of the Treasury explicitly suggested use of "widespread propaganda" to convince the public to forgo their "needless pleasures" (U.S. Treasury Department 1918: 2). The Treasury Department implemented what it called a "campaign of education" regarding the income tax. Its "essential features" included government-supplied news stories and editorials as well as encouragement of special cartoons and films. Perhaps its most intriguing feature, however, was its use of the clergy. The commissioner of Internal Revenue reported that "Thousands of clergymen, *at the suggestion of the Bureau*, made taxation the subject of at least one sermon." As a result of the "patriotic response" aroused, "dissatisfaction and complaint over the burden imposed by taxation were minimized." Government officials commented that "the groundwork was laid for securing in ensuing years prompt and regular response to revenue demands." To perpetuate its success, the Bureau of Internal Revenue advocated "the most intensive cultivation of intelligent public opinion" (U.S. Treasury Department 1919: 964–65, 974; see also Higgs 1987: 133–34).

In the second aforementioned study, Baack and Ray examined an earlier period of tax history to discover why it was that, although the 1894 income-tax statute was declared unconstitutional by the Supreme Court in 1895, a constitutional amendment was not introduced in Congress until 1909. Their results suggested the "pivotal role of federal transfer payments in securing passage of the Sixteenth Amendment in 1913" (Baack and Ray 1985a: 607). Between 1895 and 1909, government officials—acting through the secretary of War, the secretary of the Navy, and the commissioner of pensions—channeled disproportionate government military-related outlays to the states whose congressional delegations up to that time had consistently opposed the income tax. For instance, 74.7 percent of the increases in annual War Department expenditures on army arsenals, posts, and public works between 1897 and 1908 went to the 17 states that previously had

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

opposed income tax legislation. To Baack and Ray (1985b: 128–31), this and related evidence appeared "consistent with the possibility that naval expenditures and veterans benefits were used to buy state votes to support the income tax amendment." These targeted outlays and the implicit possibility of their withdrawal clearly raised the opportunity costs to affected legislators and their constituents of continuing to resist the income tax. Deliberate choices by government officials again reshaped political transaction costs influencing the role and scope of government.

Even if one grants the prevalence of political transaction-cost augmentation, still one may ask what difference it makes. Can we not expect the public, in a representative democracy, to reverse political arrangements incompatible with their preferences? Unfortunately, we cannot. Transaction-cost-increasing measures by definition raise the costs to individuals of particular forms of political action. For any given distribution of political opinion, such measures therefore drive a wedge between people's preferences and the political expression of those preferences. Policies unwanted by the bulk of the citizenry may survive.

Moreover, transaction-cost-increasing measures often alter the institutions of government, changing society's institutional bedrock. In the long run, such institutional changes tend to reshape predominant societal ideologies so as to validate existing government authority—in effect molding people's beliefs to conform to the new institutional status quo. In his study of crisis-induced changes in government authority, Higgs (1985, 1987) has shown how institutional change generates self-validating ideological change as people become accustomed to and "learn to like" the new institutional arrangements. Such ideological changes also occur as a result of people's reticence to express unpopular views in public. As Timur Kuran (1987, 1991, 1993) has shown, given the institutional status quo, people often have strong incentives to misrepresent or "falsify" their preferences in public discourse. As a result, succeeding generations receive less exposure to public discourse questioning the status quo, and more exposure to public discourse affirming it. Accordingly, subsequent generations may perceive many societal decisions embodied in the institutional status quo as settled, despite widespread (though unexpressed) private preferences to the contrary. Over successive generations, institution-validating ideological change is the likely result.

Political transaction-cost manipulation thus matters in a larger sense because it artificially redirects political action, facilitating institutional changes that ultimately distort public discourse and channel ideological change in ways potentially inconsistent with people's initial prefer-

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ences—and, sometimes, preexisting constitutional law. For theoretical reasons explored in greater detail elsewhere, this politico-economic sequence typically eventuates in greater dependence on government (Twight 1993).[7] The balance of this paper analyzes the extent to which the model is consistent with the establishment and expansion of income tax withholding in the United States.

Early History of Withholding: A Reinterpretation

> Wherever an income tax has been in practice for any time the small incomes as well as the large are taxed; and it is the small incomes which yield the largest revenue to the state.
>
> —Treasury official Worthington C. Ford (U.S. Senate 1894)

U.S. Income Tax and Withholding Experience Before 1940

An income tax was first employed in the United States during the Civil War. Although many, including the secretary of the Treasury, desired longer retention of the Civil War income taxes, the taxes were widely viewed as emergency measures and were repealed in 1872. This was a time when even the commissioner of Internal Revenue recommended repeal of the income tax, writing to the chairman of the House Ways and Means Committee that he regarded the income tax as "the one of all others most obnoxious to the genius of our people, being inquisitorial in its nature, and dragging into public view an exposition of the most private pecuniary affairs of the citizen" (U.S. House 1871: 1). Such opinions provide a baseline against which to assess later changes in public sentiment.

Though proposed many times, income tax legislation was not enacted again until 1894. Consistent with a transaction-cost-manipulation model, Congress labeled the 1894 law "An act to *reduce taxation,* to provide revenue for the government, and for other purposes." When challenged in the case of *Pollock v. Farmers' Loan and Trust Company*, the income tax law was held unconstitutional by the U.S. Supreme Court because it established a "direct" tax on real property and invested personal property deemed unconstitutional without apportionment among the states according to population as mandated by the Constitution (157 U.S. 429, 158 U.S. 601, 1895).[8]

Income taxes temporarily were stymied. There was strong sentiment in the Senate to pass similar legislation and again confront the Supreme Court on this issue. Wanting to avoid such a confrontation, President Taft in 1909 recommended both a corporate income tax, labeled as

[7]The emergence and entrenchment of Social Security legislation in the United States has proved consistent with this model's description of the nexus between transaction-cost manipulation, institutional change, and ideological change (see Twight 1993).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

an "excise" tax to avoid constitutional censure, and a constitutional amendment authorizing taxation of income from all sources without apportionment among the states (U.S. Senate 1909). Many staunch opponents of income taxation nonetheless supported Taft's proposal, hoping that the corporation income tax and the cumbersome amendment process would erode support for more broadly based income taxation. Congress submitted the proposed Sixteenth Amendment to the states for ratification in 1909.

A confluence of circumstances facilitated adoption of the income tax amendment. Chief among them was widespread belief that the existing federal tax system, with its reliance on tariffs and excise taxes, unfairly burdened the less affluent. Noting "a growing conviction among people from all walks of life that the existing tax system failed to reach the great fortunes that had been amassed as a result of industrialization," John Buenker (1981: 185) identified such beliefs as the "single most important reason for the eventual enactment of the federal income tax." Then as now, people's tax preferences often were driven by beliefs about tax incidence. Detailed studies of the history and politics of the period indicate intense desire on the part of various regional and economic groups to rearrange taxes to make others pay a disproportionately high share of governmental costs (Buenker 1981, Ekirch Jr. 1981). Thus most low-income Southern and Western states endorsed federal taxation based on "ability to pay" and favored a graduated federal income tax differentially burdensome to wealthier states in the North and East. As noted in the preceding section, the apparent manipulation of federal transfer payments also may have contributed to some states' approval of the amendment (Baack and Ray 1985a, 1985b). Widespread concern about cost-of-living increases partially attributed to import tariffs, along with increases in U.S. exports and military expenditures, created additional pressures to find alternative sources of federal revenue. Strengthened by elections in 1910 that reduced Republican representation in many state legislatures, these mutually reinforcing conditions led many states previously opposed to the income tax to favor the amendment.

As the Sixteenth Amendment moved toward ratification, some state governors waxed eloquent in their support of the income tax amend-

[8]The Supreme Court did not comment on the law's taxation of gains from business and employment, citing "the instances in which taxation on business, privileges, or employments has assumed the guise of an excise tax [not subject to apportionment] and been sustained as such" (158 U.S. 635). Later writers and judges interpreted *Pollock* to mean that the validity of such taxation was recognized and that there was "no dispute" about that issue (*Brushaber v. Union Pacific Railroad Co.*, 240 U.S. 1, 17, 1915). For extended discussion of the *Pollock* case, see Higgs (1987: 99–103), and Arthur Ekirch Jr. (1981: 168–71).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ment. Governor John Franklin Fort assured the people of New Jersey that the citizenry can be relied upon "to see that their representatives make no unjust exactions in the way of taxation or in the curtailing of the rights of the States or otherwise," that the amendment "is vital to the safety and security of the Republic," and that it "is without danger in the power conferred" (U.S. Senate Doc. No. 365, 1910: 5). Some believed that income taxes authorized by the amendment would be implemented only during emergencies. Presaging later transaction-cost augmentation, Senator Norris Brown (R., Nebraska) asserted that the income tax amendment "lays no tax, promises to lay none, but simply and solely restores to the people a power many times sustained but finally denied by the courts" (U.S. Senate Doc. No. 705, 1910: 6).

The 16th amendment became constitutional law in February 1913. Contrary to Senator Brown's implication, income tax legislation was adopted in October of that year.

The 1913 statute authorized withholding of income taxes "at the source"—that is, extraction of income taxes from taxpayers' pay envelopes before salaries were paid. Precedent existed in the income tax withholding for government employees during the Civil War (Bopeley 1943). However, the 1913 law's withholding provision proved to be a great irritation to taxpayers, a fact downplayed in later discussions of withholding. Based on public criticism, Treasury Secretary William G. McAdoo reported that "it would be very advantageous to . . . do away with the withholding of income tax at the source" because it would "eliminate a great deal of criticism which has been directed against the law" (U.S. Treasury Department 1916: 19). The following year the commissioner of Internal Revenue, in a report also signed by McAdoo, formally recommended that "the provisions of law requiring the withholding of the normal income tax at the source of the income be repealed" (U.S. Treasury Department 1917: 674). The authority for withholding was withdrawn in 1917, not to be resurrected until the 1940s.

The Current Tax Payment Act of 1943

Despite the 1913–16 experience, Congress in 1943 passed the Current Tax Payment Act, establishing the broad-based income tax withholding that has continued to this day. The important politico-economic question is how and why. This section discusses transaction-cost-increasing strategies used to structure political support for a policy previously so unpopular with the public. We will contrast the ostensible and actual purposes of the withholding law, analyzing the political mechanisms that made its passage possible.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Conventional wisdom suggests that withholding became advantageous to the public with the vast expansion of income taxation that occurred during World War II. In fact, the military crisis facilitated establishment of institutional mechanisms that served long-run interests of government and its functionaries rather than the public, with crisis providing an essential ingredient and cover for all manner of misrepresentations used to secure passage of the withholding act. As Higgs (1987), Forsythe (1977), and others have noted, real or purported crisis often provides a carte blanche for expansion of government authority. In the more general framework employed here, crisis facilitates transaction-cost augmentation by influencing its determinants—providing an appealing rationale for transaction-cost-increasing measures, stimulating executive and party support for such measures, prompting favorable media coverage, and shortening the public's time horizon so as to focus attention on the emergency at hand and deflect attention from transaction-cost-increasing features of proposed legislation.

We know that World War II prompted transformation of a tax long endorsed by the public as a tax on the rich into a tax on the masses—a "people's tax" in the familiar words of Treasury Secretary Henry Morgenthau Jr. (U.S. Senate Hearings July-August 1942: 3). The numbers have been widely reported elsewhere. A Treasury Department official testified in early 1943 (U.S. House Hearings 1943: 2):

> Up until 1941 we never received as many as 8,000,000 individual income-tax returns in a year. In 1941 that number increased to 15,000,000; in 1942 it increased to 16,000,000. This year we expect 35,000,000 taxable individual income-tax returns.

It was one thing to pass the laws that authorized such taxation. The troubling question for government officials was how to assure that the taxes would be paid. Early on, they recognized that income tax withholding could get the job done; the problem was how to sell it to a public previously hostile to such measures.

Ostensible versus Actual Purposes of Withholding. In 1941 Albert G. Hart, professor of economics at Iowa State College, proposed a general plan for collection of income taxes at the source (U.S. House Hearings 1941: 330–48). The next year Treasury Secretary Morgenthau (U.S. House Hearings 1942, vol. 1: 5) recommended income tax withholding, presenting it as a "more convenient method for the payment of income taxes." Government concern for the well-being of the taxpayer was the dominant theme. Throughout this period the Treasury Department consistently portrayed the withholding proposal as providing taxpayers "a way of meeting their tax obligations with a

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

maximum of convenience and a minimum of hardship" (U.S. House Hearings 1943: 9). As Treasury official Randolph Paul (U.S. House Hearings 1943: 10) put it:

> The tax has been broadened to reach many millions of additional taxpayers with small incomes and little experience in planning their finances to meet large bills at infrequent intervals. . . . A suitable pay-as-you-go method will be of great assistance to millions of persons.

The fact that withholding had been tried before, and that the public had strongly opposed the earlier withholding system, seldom was mentioned.[9]

As the president and Congress imposed ever higher income taxes, tax payment was wrapped in patriotism. In congressional hearings as in government propaganda efforts documented by Jones (1989), sacrifice was a dominant theme. Treasury officials (U.S. House Hearings 1941: 49; Higgs 1987: 202–03) labeled proposed tax increases "light indeed as compared to the sacrifices which large numbers are undergoing in entering military services."[10] Secretary Morgenthau (U.S. Senate Hearings July-August 1942: 8) urged Congress to adopt a "courageous tax bill," avowing that "acceptance of sacrifice on the home front is a yardstick of our determination to win the war."

Although taxpayer convenience and patriotic sacrifice were the avowed purposes of income tax withholding, the actual objectives—though not trumpeted to the public—were candidly acknowledged in congressional hearings. These harsher objectives included increasing government revenue, enforcing payment of taxes, and muting taxpayer resistance. Treasury officials viewed pay-as-you-go withholding as a way to "collect some money from people who would not otherwise make any report on income," testifying that "We cannot get those fellows unless we have the collection-at-the-source method" (U.S. Senate Hearings July-August 1942: 137). They advocated "us[ing] the tax system as we would a delicate surgical tool" (Paul 1943: 327). A recurrent theme was "the far greater collectibility of the tax if it is collected currently" (U.S. House Hearings 1943: 76).

Fear of taxpayer resistance was prevalent. One witness (U.S. House Hearings 1943: 391) warned that, without withholding, "taxpayers will simply throw up their hands and in a defiant tone say, 'Try and collect'." That fear surfaced again in an exchange (U.S. Senate Subcomm. Hearings, August 19, 1942: 61) regarding withholding between Senator Bennett Champ Clark (D., Missouri) and Treasury's Randolph Paul:

[9]I found only four brief references to it in thousands of pages of hearings.

[10]On the widespread use of this rationale during World War II, see Higgs (1987).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

> Senator Clark: Psychology almost certainly ought to be considered in the tax year. Some British Chancellor of the Exchequer once said: 'Taxation consists of getting the greatest amount of money with the least amount of squawks.'
>
> Mr. Paul: Do you think if we cut down the squawking under this method we could raise the individual tax rates?
>
> Senator Clark: That is what I am trying to find out: How we can raise the greatest amount of money with the least amount of hardship on the taxpayer.

As transaction-cost-augmentation theory suggests, "squawking"—vocal resistance to taxation—was viewed as manipulable, controllable by officeholders' deliberate decisions to change institutional mechanisms of government.

Long-term advantages of withholding to the government were apparent to Congress. As Representative Donald H. McLean (R., N.J.) put it (U.S. House Hearings 1943: 85), the advantages involved "protecting the Government revenues not only now, but for all times to come." McLean believed that everyone felt "the need for the change in the collection method, due to the increase of the number and type of taxpayers that we have brought into the system." Witnesses testified (U.S. House Hearings 1943: 187) that it would be "good business" for the government: government "will have more revenue; . . . its people will pay better and be happier about it."

Nonetheless, an effort was made to maintain a facade of solicitous concern for the taxpayer. Whenever a crack appeared in the facade it was quickly smoothed over—as when a Treasury official discussing withholding (U.S. House Hearings 1943: 32–33) referred to the "person *against whom* the method was applied" and quickly corrected himself to say "or I might say in whose *favor* it was applied."

Political Strategies for Effectuating Withholding. The key strategies used to obtain support for income tax withholding in 1943 all entailed political transaction-cost augmentation. Government officials artfully employed national defense language, tax-cost information, and promises of "tax forgiveness" to engineer support for a withholding system at root designed to enhance and protect government revenue for all times to come. The above-noted conflict between the government's actual objectives and its publicly promoted objectives formed only one part of a systematic pattern of transaction-cost manipulation documented below.

Disingenuous use of the defense theme to secure tax increases was acknowledged in congressional hearings. Representative Frank Carlson (R., Kans.), admonishing witnesses to use such language, reminded them that the House Ways and Means Committee "passed

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

a 10-percent increase in our income and corporate taxes a year ago by calling it a *defense* tax." He opined (U.S. House Hearings 1942, vol. 1: 508) that "the suggestion that we call this tax a war tax is a good one." The power of the war image to overcome political resistance also was evident in polling data to be discussed below. Similarly, in discussing the issue of "forced savings," Representative A. Willis Robertson (D., Va.) noted that the "word 'forced' is not a euphonious name" and that it "would be much better if we should call it 'Victory savings,' or something of that kind" (U.S. House Hearings 1942, vol. 1: 108). Treasury official Randolph Paul agreed. The language enwrapping revenue legislation was not lightly chosen.

Other forms of political transaction-cost manipulation also proved instrumental in securing passage of income tax withholding legislation. Those strategies were evident in government officials' handling of (1) the present-value issue, (2) the Ruml plan, and (3) the final debates.

1. Present-value issues were pivotal to important misrepresentations surrounding income tax withholding proposals in the early 1940s. In the 1920s and 1930s, income taxes had been due and payable on March 15 following the end of the tax year—for example, 1938 taxes were due on March 15, 1939, and could be paid either in one lump sum on that date or in quarterly installments during 1939. The proposed system would require employers to extract tax payments out of each paycheck during the tax year, so that a given year's taxes would be paid largely during that same year.

Treasury officials repeatedly testified to Congress that such withholding of income taxes—current collection at the source—represented "no additional tax." On dozens of occasions, Treasury official Randolph Paul and other government spokesmen testified:

> This collection at the source mechanism is nothing but a mechanism for collection. It is not an additional tax. . . . It merely speeds up the collection (U.S. House Hearings 1942, vol. 1: 100).
>
> It should be kept in mind that collection at the source does not in itself increase or decrease the tax liability of the taxpayer (U.S. House Hearings 1943: 11).

Given the expert witnesses' knowledge of present value, statements so seriously misleading to Congress and the public could not have been inadvertent.

Treasury officials and members of Congress who repeated these statements implicitly treated dollars today as identical in value to future dollars. This indefensible foundation of the Treasury's analysis was not made clear to Congress or the public. Indeed, in his congressional testimony, Randolph Paul simply added up an individual's tax

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

liabilities over various years, without making a present-value computation, to compare that person's total tax burden under various proposals (U.S. House Hearings 1943: 23 ff.). When members of Congress probed too closely, Paul and other officials usually sidestepped their questions.

Nonetheless, some astonishing statements were elicited. Consider the 1942–43 House hearings on this issue. When Representative Thomas A. Jenkins (R., Ohio) protested, "I have seen taxes collected after they have accrued, but I never saw them collected 6 months ahead of time," Treasury Secretary Morgenthau replied, "You are putting it very bluntly, but that is what we are proposing to do." The Treasury Department repeatedly acknowledged that this represented "payment in advance" (U.S. House Hearings 1942, vol. 1: 22, 57, 78). Yet Treasury officials insisted (U.S. House Hearings 1943: 36):

> There is nothing in collection at the source that imposes any additional tax burden. Collection at the source relates entirely to the method and time of payment. It advances payment which otherwise would not be made until the following year, under our present system, to the current year—indeed, to the very time when the payment to the salary recipient is made.

Whether or not members of Congress understood the concept of present value, it is clear that Treasury officials did. Milton Friedman, then working for the Treasury Department, certainly was cognizant of present values when he stated (U.S. Senate Subcomm. Hearings, 19 August 1942: 58) to a congressional subcommittee evaluating alternative tax plans, "You must also take into account the timing of the receipts." Randolph Paul alluded to the government's "power to make up the loss [associated with eliminating certain tax liabilities] by *compelling quicker collections*" (U.S. House Hearings 1943: 17). Treasury officials further demonstrated their understanding of the time-value of money by recommending that the Bureau of Internal Revenue be required to pay interest on amounts refunded under the new tax law (U.S. Senate Hearings 1943: 35).

Moreover, before withholding was reestablished in 1943, the government sold interest-bearing "tax anticipation notes" which private citizens could buy during the year to generate interest to help pay their taxes when they were due the following year (see *U.S. Cong. Rec.-Senate* 14 May 1943: 4419). Like other investment vehicles, such tax anticipation notes enabled taxpayers to set aside a smaller amount in the present to satisfy any given future tax liability. In contrast, under the proposed withholding system—with identical tax rates—the taxpayer would have to forgo a larger sum in present-value terms to satisfy the tax collector. The government, not the taxpayer, would

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

receive the benefits obtainable from earlier command over that income.

Nonetheless, the Treasury Department's claim that withholding was not an additional tax was repeated by members of Congress on the House and Senate floor and elsewhere. On this fundamental issue, government officials systematically raised the transaction costs to the public of assessing the proposals at hand. Accordingly, while other features of the bill prompted bitter dispute, by the time the Current Tax Payment Act reached the floor of Congress there was no dispute about current withholding of income taxes at the source. Ironically, transaction-cost augmentation was employed to curry support for a proposal that, once adopted, in turn would serve as a key mechanism for increasing other political transaction costs facing the public.

2. Transaction-cost augmentation also took other forms, including a "paper forgiveness" of income taxes that came to be known as the Ruml plan. As Congress considered various withholding proposals, a key transitional problem became apparent. Immediate conversion to a pay-as-you-go system seemed to entail double taxation in the transition year. That is, if a pay-as-you-go system were adopted in 1943, during 1943 people would be required to pay both their 1942 taxes (under the old law) and their 1943 taxes (via the new withholding arrangement). Although Treasury officials thought that was a fine idea, most others disagreed.

Accordingly, various proposals aimed to soften this effect. The Treasury was willing to spread out the extra year's tax over an extended period to accomplish the transition. However, the idea that captured the public's attention was Beardsley Ruml's proposal (first made in the summer of 1942) to cancel or "forgive" one year's tax, treating amounts paid or withheld in 1943 as payments toward a person's 1943 tax and eliminating 1942 tax liability.[11] Using the metaphor of daylight savings time, Ruml proposed to set the "tax clock" ahead one year.

Two things stand out from the convoluted history of Ruml's proposal. The first is that it was absolutely critical to—and perhaps the proximate cause of—public acceptance of income tax withholding in 1943. The second is that the tax "cancellation" involved was a sham and was understood to be a sham by a significant number of government officials involved in its passage. Both of these conclusions point to the transaction-cost-increasing role of the Ruml plan in securing passage of the Current Tax Payment Act of 1943. Sham or not, taxpayers liked the sound of the words, and government officials were

[11]Beardsley Ruml, of R. H. Macy & Co., was at the time chairman of the Federal Reserve Bank of New York.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

attentive to the nuances. The psychology of taxation was a recurrent theme. As Ruml testified, "there is a power in words to evoke emotion, and double taxation evokes emotion." He explained (U.S. Senate Subcomm. Hearings 19 August 1942: 5) that "People don't believe in double taxation, even though the single taxation may be the sum of the two." The Current Tax Payment Act played on this psychology.

The polling data support the supposition of Representative Robert L. Doughton (D., N.Car.) that "one of the principal reasons for the popularity of this plan [is] the fact that it relieves the taxpayers of the year's taxes."[12] Although pay-as-you-go income tax withholding had been under discussion in Congress since 1941, by June 1942 public sentiment remained quite equally divided on the idea. For example, asked in May and June 1942 if they would "like to have a regular amount deducted from each pay check" to pay their federal income tax, 43 percent of the respondents said no, 50 percent said yes, and 7 percent were undecided. In a similar poll conducted on February 3, 1942, 45 percent said no, 45 percent said yes, and 10 percent were undecided (Gallup 1972: 338; Cantril 1951: 324). It was not a groundswell. The only polls during this pre-Ruml-plan period that found substantial support for withholding were those that inserted the phrase "to help the war effort" in their question.[13]

However, after the Ruml plan was introduced in July-August 1942 and the idea of tax cancellation or forgiveness was touted in the popular press, there was a dramatic change in public sentiment. Polls conducted November 1942 to April 1943 found that a steadily rising percentage of respondents favored pay-as-you-go withholding, with support for the proposal ranging from 65 percent to 79 percent without mentioning war in the question. A similarly large percentage of respondents reported familiarity with the Ruml plan. Of the 81 percent of respondents who had heard of the Ruml plan in January 1943, 90 percent of those who expressed an opinion about it favored the plan. In February 1943, 42 percent of respondents expressed their belief that the Ruml plan would mean that they would not have to pay tax on their 1942 incomes (Cantril 1951: 324–25; Gallup 1972: 366, 371).

Thus it appears that the public jumped at the bait of tax forgiveness. The issue is whether the hoped-for cancellation was real. On the most obvious level, collecting two years' taxes in one year in the process

[12]U.S. House Hearings 1943: 184. Chairman Doughton incorporated this phrasing in a rhetorical question.

[13]Two poll questions in May 1942 that described withholding as an idea "to help the war effort" found 64 percent and 72 percent of respondents supported the idea (Cantril 1951: 324).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

of moving to a current collection basis would seem to imply greater tax revenues for the government, suggesting that cancellation of 1942's tax would reduce tax revenues. Therefore, one question is whether the government's revenue actually was expected to fall if a comparison was made between double taxation and tax forgiveness in the transition year (i.e., a comparison of current collection at the source with and without tax forgiveness). But that is not the only comparison to be made. Taxpayers naturally are concerned with how they will fare under the existing system versus a proposed new system. Thus another critical comparison is the government's revenue "take" without withholding (the old law), and the government's take with tax forgiveness coupled with current collection at the source (a Ruml-like new law).

Despite their opposition to tax cancellation, Treasury officials did not expect tax revenues to fall viewed from either of these perspectives. Faced with rising popular support of a Ruml-type plan, they repeatedly acknowledged that the post-cancellation situation could be expected to entail a greater tax-take for the government—whether compared with the old law or with a hypothetical situation involving two years' taxes in one. For instance, Randolph Paul testified regarding tax cancellation:

> The Government by forgiving a year's tax liabilities would be discarding assets. . . . The Government differs from the business in that it has the power to make up the loss by compelling quicker collections and by imposing additional taxes on the same or other people . . . the cash receipts of the Treasury could be maintained even though the tax liability was forgiven (U.S. House Hearings 1943: 17).

Moreover, Paul acknowledged that, in the expected environment of rising incomes, the government's tax-take would increase under a Ruml-type plan compared with its revenues under the old law, even if tax rates were not raised. With present-value issues just beneath the surface, Paul testified (U.S. House Hearings 1943: 18):

> Each individual subject to taxation in 1942 has 1 year's liability canceled, but he is at the same time required to pay another year's liability sooner than he otherwise would. Individuals who were not taxpayers in 1942, but who become taxpayers subsequently, will be obliged to pay their liabilities 1 year sooner than under existing law. Individuals who die, or who cease receiving an income, pay the Government 1 year's less taxes, but by and large the money loss on their account is offset by the gain from new taxpayers who begin paying their taxes a year earlier. . . . The payments dropped out will be spread over a period of years. If any given year is a year of higher national income . . . the actual receipts of the Government for that span of years would be increased by the change.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Former Treasury Department official Elisha Friedman openly called it a "paper forgiveness." Referring to lower-income taxpayers as "little people," he stated (U.S. House Hearings 1943: 503) that he "would agree to 100 percent forgiveness for little people, because, frankly, it is a paper forgiveness." Noting that withholding at the source "makes possible higher tax rates than under the present method," he testified (U.S. House Hearings 1943: 491, 492, 503):

> The 'forgiveness' of the small brackets is merely temporary. . . . They will pay more later. . . .You will forgive the 1942 tax for the little people but in 1944 and 1945 they will be paying at a higher rate. . . . Ours is a paper forgiveness for the low brackets.

Some in Congress resisted the language of tax cancellation. Condemning "legislative legerdemain in the cancellation of 1942 income tax liabilities," Senator Robert M. LaFollette Jr. (R./Progressive, Wisc.) expressed his belief that, in the context of rising war expenditures, the "average taxpayer would rather learn the bad news now . . . than be misled by the false sound of cancellation" (U.S. Senate Report No. 221, 1943, part 2: 1). As Senator Henry Cabot Lodge Jr. (R., Mass.) put it to a Treasury witness (U.S. Senate Hearings 1943: 94), "if you go at it from the standpoint that you live by, that you feed your children on, those things, there is no cancellation at all, is there?"

3. In examining the final congressional debates on this bill, it is not my intention to chronicle either the procedural maneuvers or the detailed differences between contending bills as income tax withholding legislation moved toward passage. Excellent summaries already exist (U.S. Senate 1946). My focus is on central politico-economic strategies that facilitated passage of such legislation.

The conference bill ultimately passed by the House and Senate involved compromise on everything except the fundamental idea of withholding at the source. As Randolph Paul had stated (U.S. Senate Hearings 1943: 2), the three leading bills "reflect[ed] essential agreement on the major issue of current payment." On the Senate floor, Senator Arthur H. Vandenberg (R., Mich.) reiterated (*U.S. Cong. Rec.-Senate* 2 June 1943: 5209) that "No one questioned at any turn of the road the desirability and necessity of having collection at the source and making the Nation current with its taxes."

The only disagreement concerned the degree of tax cancellation for 1942. President Roosevelt, having called for a $16 billion tax increase, stated in writing to the chairman of the House Ways and Means Committee and the Senate Finance Committee that he would veto legislation authorizing 100 percent cancellation of 1942 tax liability. Another constraint was concern that the rich would benefit more

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

than the poor from tax cancellation. In particular, many wanted to avoid abating taxes on the "windfall profits" of war contractors.

With certain qualifications and exceptions, the bill finally passed authorized 75 percent cancellation of one year's tax liability. Two windfall profit provisions were included in the conference committee bill. In general, the final bill required payment of the higher of one's 1942 and 1943 tax liability. If someone died in 1943, or had much lower income in 1943 than in 1942, that person would not thereby avoid his 1942 tax liability. Moreover, to prevent recipients of "war profits" from receiving a boon, the bill set an additional cap on tax forgiveness for those whose lowest income in 1942–43 exceeded their income in a selected base year (1937, 1938, 1939, or 1940) by more than $20,000 (U.S. Senate 1946).

Widespread awareness of the transaction-cost-increasing features of the Current Tax Payment Act was evident in the final debates. Despite allusions to alleged mutuality of interest, it was widely understood that income tax withholding was chiefly in the interest of the government, not the taxpayer. Calling current collection "the crux of the whole matter," Senator William W. Barbour (R., N.J.) told the Senate (*U.S. Cong. Rec.-Senate* 12 May 1943: 4271) that "the best interests of the Government will be served if the new tax law requires that taxes be paid while the taxpayer has the money to pay them." Senator Harry Flood Byrd (D., Va.) said (*U.S. Cong. Rec.-Senate* 13 May 1943: 4336) that it was "of great interest and importance to the Treasury, as well as the Government as a whole, that taxes be placed on a pay-as-you-earn basis." Senator David I. Walsh (D., Mass.) added that withholding "is of more benefit to the Treasurer than to anyone else" and "means that the Treasury will be able to collect future taxes" (*U.S. Cong. Rec.-Senate* 2 June 1943: 5210).

Similarly, there was no doubt in the minds of many representatives that the result of withholding, even with tax forgiveness, would be an increase in the tax burden on the public. Although Senator Walter F. George (D., Ga.) as chairman of the Finance Committee repeated the official line that the withholding bill "does not deal with rates directly, nor does it affect the burden imposed under varying rates upon the taxpayers," others were more candid. Senator Barbour noted that "the change in the method of tax collection will unquestionably increase the flow of revenue to the Treasury." Reinforcing this point, Senator John A. Danaher (R., Conn.) observed (*U.S. Cong. Rec.-Senate* 12 May 1943: 4268, 4272, 4282) that "The fact of the matter is the Treasury collections will go up annually rather than down." Senator Byrd predicted that "before the ink is dry on the signatures" establishing a Ruml-type bill as law, the Treasury "will call upon

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the Congress to increase the existing tax rates in proportion to the cancelation [*sic*] and forgiveness we extend to the taxpayers." He believed that "so-called benefits to the taxpayer" would then "quickly sink into complete oblivion" so that "most taxpayers would be injured rather than benefited" (*U.S. Cong. Rec.-Senate* 13 May 1943: 4337). Advocates of a Ruml-type plan openly boasted (*U.S. Cong. Rec.-House* 3 May 1943: 3841) that it "would actually bring in $3,000,000,000 more revenue to the Treasury this year than would the present law."

The fact that tax forgiveness was both a sham and an essential ingredient of public support for the income tax withholding bill was widely discussed in the final debates. Senator Tom Connally (D., Texas), an opponent of the Ruml plan who believed it portended a loss to the Treasury, asserted that the bill "is really intended to fool people." He believed that the Ruml plan would be "blown out of the water" by those "whooping up the Ruml plan" if they became convinced that they were "not going to get any money back" (*U.S. Cong. Rec.-Senate* 14 May 1943: 4408). Senator Clark of Missouri stated (*U.S. Cong. Rec.-Senate* 12 May 1943: 4275) that he "never believed that there was any forgiveness or any personal advantage to anybody in the [proposed] system," perceiving "great governmental advantage in having everyone current with his taxes" and enabling government to "collect the taxes as the taxpayer earns them."

Nonetheless, while some congressmen understood the issue, others succumbed to its apparent complexity. Complexity here facilitated the adoption of transaction-cost-increasing measures, allowing the "experts" to steer the outcome to suit their own interests. Senator Connally of Texas stated that he did "not think there is anyone on the [Senate Finance] committee who completely understands all the angles," noting that taxation had become so complex that the Finance Committee "could never make any progress or headway if it did not have available the experts of the Treasury." He described (*U.S. Cong. Rec.-Senate* 14 May 1943: 4409) the relation between the committee and the Treasury experts as follows:

> When a question arises we call on them for information as to what the effect of certain proposals would be, what the repercussions would be, what the reactions would be, and we are obliged to act on the basis of the information thus furnished.

Consistent with transaction-cost-augmentation theory, complexity not only encouraged reliance on experts but also provided political cover for those who took their advice.

Crisis also facilitated passage of income tax withholding legislation. In the dispute over forgiveness or cancellation of 1942 taxes, outraged

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

opponents of any cancellation impugned the patriotism of their adversaries and asked how one could in good conscience cancel taxes when U.S. soldiers were dying in battle. Countless allusions to "our men and boys . . . dying to win victory and save our country" peppered the debates. Compounding the difficulty of understanding the actual import of the Ruml plan, those convinced that it signified reduction in government revenues invoked the "price in life and limb" being paid on the battlefield, stating that "no sacrifice however great of the citizen taxpayer at home can compare with the privations of the soldier in the field" (*U.S. Cong. Rec.-House* 4 May 1943: 3923, 3928). War and the oft-expressed desire to limit inflation by absorbing citizens' spending power provided appealing rationales conducive to approval of the income tax withholding measure.

Thus the Current Tax Payment Act of 1943 became law, both product and instrument of transaction-cost augmentation. Though the Act was widely supported by the citizenry, we have seen that wartime public support rested on misunderstandings actively encouraged by government officials. However, the Current Tax Payment Act of 1943 did not arise in an institutional vacuum, nor did its support erode as the public learned more about its effects. That deception was only one part of the fabric of transaction-cost augmentation surrounding this issue is shown below as we consider the interplay between transaction-cost manipulation, institutions, and ideology in shaping the evolution of U.S. income tax withholding.

Changing Institutional and Ideological Contexts: The Role of the Status Quo

> Taxes which are easy to collect tend to be extended and expanded with similar ease by legislative bodies. The withholding provisions make it easy for the Treasury to collect taxes from wage earners and low-income groups. We must be ever vigilant to prevent this ease of collection from being used as a lever further to lower personal income tax exemptions or otherwise to impose new burdens on low-income groups.
>
> —National Lawyers Guild
> (U.S. House Hearings 1942, vol. 2: 2302)

The institutional and ideological status quo provides context for, and sets constraints upon, further changes in institutions and ideologies. Politico-economic developments are path-dependent.[14] In light of the preceding section's evidence of government officials' use of transac-

[14]See Higgs (1985: 2–3 ff.; 1987: 57–74).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

tion-cost augmentation in effecting adoption of the Current Tax Payment Act of 1943, this section examines the broader institutional context that shaped this result and the long-run institutional and ideological changes that followed it.

The theory discussed earlier in this article predicted that, when institutional change occurs as a result of transaction-cost augmentation, the long-run outcome is likely to be further authority-legitimating institutional and ideological change.[15] In the case of U.S. income tax withholding, key evidence is to be found in the institutional structures out of which it arose and into which it developed. We will see below that government officials themselves viewed the relevant institutional changes as parts of an incremental process, working to strengthen the "machinery of taxation" over time.

Institutional Precursors of the 1943 Act: Building a "Tax Machine"

Against the backdrop of the failed experiment with income tax withholding during the 1913–16 period, two institutional changes had occurred that significantly influenced the political viability of the 1943 legislation. First, the Social Security Act was adopted in 1935. While extensive manipulation of political transaction costs characterized adoption and expansion of that law (Twight 1993), most important here is that the social security law was funded by means of a payroll tax withheld at the source (Leff 1983). This funding mechanism emerged in the context of a law widely but falsely promoted as giving each "contributor" an "account" in Washington, D.C., that would provide income security in his old age.

Second, in 1942 Congress and the president established a so-called "Victory Tax" over and above other income taxes. Congressional attention to euphonious labeling carried the day. This Victory Tax differed from other income taxes in that it entailed a flat-rate tax of 5 percent above a $624 exemption and was required to be withheld at the source by employers.

These two taxes undergirded the 1943 withholding law. Members of Congress and witnesses in congressional hearings repeatedly called attention to the linkage. At the outset, in proposing broad-based withholding (U.S. House Hearings 1941: 345), Professor Albert G. Hart reminded key congressional committees:

> We are already collecting taxes, or 'contributions' if you like, from a large part of our wage earners and salaried people under the Social Security Act. That offers a nucleus for this reorganization.

[15]See notes 1–7, *supra*, and accompanying text.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

> Besides this, we have already a system of reporting at the source by employers, a force of internal revenue field agents, and so forth. Accordingly, *we have the makings of an adequate tax machine*. Most of the parts are there.

Treasury official Randolph Paul testified (U.S. House Hearings 1943: 12, 78) that the "essential machinery" for collection at the source already was established under the Victory Tax, and that the "social security tax has provided a basis of experience on which we have had to draw." Treasury Secretary Morgenthau (U.S. Treasury Dept. 1944: 108) described the Victory tax as "a proving ground for the withholding principle." A Senate report (U.S. Senate Report No. 221, 1943 part 2: 17) noted that the methods of collection mandated by the proposed legislation "have been coordinated generally with those applicable to the Social Security tax . . . to facilitate the work of both the Government and the employer." The conference committee report on the 1943 bill described it as a change "to a system of collection, payment, and administration based upon the principles underlying the collection of the social-security tax on wages" (U.S. House Report No. 510, 1943: 41).

Experience with the earlier laws was crucial. Stephen E. Rice, employed by the Senate's Office of the Legislative Counsel, testified (U.S. Senate Hearings July-August 1942: 136) that "All of the employers have had 7 years' experience" with the Social Security Act, and "they will be in a much better position to do this job than they were to do the social security job back in 1936 when it first went into effect."

But it was not just experience-induced ease of administration that encouraged policymakers in 1943; it was also expected diminution of public resistance born of institutional familiarity. When Representative Donald H. McLean (R., N.J.) inquired "why the compulsory payment at the source features of the 1913 act were abandoned," Treasury official Paul's response (U.S. House Hearings 1943: 82) captured the resistance-eroding effect of an institutional foot in the door:

> At that time taxes collected under an income tax system was [*sic*] something new in this country and I think it is fair to say there was some resistance to collecting at the source. . . . We were not used to being income tax payers, but now we have gone along for a period of about 30 years under the income tax system and I think the analogy is far from being very relevant.

Precisely so. Institutional change born of transaction-cost augmentation reshaped government authority in ways that, over time, engendered institution-legitimating changes in society's dominant ideologies and heightened receptivity to further authority-expanding institutional change. The institutional status quo in 1942–43—including social

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

security and "victory" withholding taxes themselves put in place through transaction-cost augmentation—set the stage for the further increment in government authority represented by the Current Tax Payment Act.

Given this sequence, the appeal of incremental change is clear to those who wish to alter fundamental institutions and ideologies in a society. As the history of U.S. social security legislation and tax policy demonstrates, incrementalism increases transaction costs of opposing institutional change due to the lower perceived marginal benefits associated with resisting piecemeal changes. What is remarkable is not that policymakers understand this but that they talk about it—at least when they are no longer in office. Elisha Friedman, an economic consultant formerly employed by the Treasury Department who was greeted by Treasury officials as a "long-lost brother," was candid. Describing how to extract the maximum out of people's pay envelopes (U.S. House Hearings 1943: 491, 500, 505), he spoke admiringly of Fraser Elliott, the Canadian Commissioner of Taxation:

> [Elliott] made it plain that an essential principle in taxation is 'Don't do anything suddenly.' . . . He said 'We must follow a policy of doing things so gradually that it is politically acceptable to the voters.' . . . You have got to get the people's minds accustomed to things. You have got to work out the political angle, and you have got to work out the administration. You cannot do it suddenly.

Elisha Friedman (U.S. House Hearings 1943: 505) recommended "continu[ing] the tax fantasy a little bit" in order to maximize extraction of resources, explaining that "If you were trying to cure a man of the drink habit, you wouldn't cut off his supply of liquor all at once. You would do it gradually." In this ex-Treasury official's view, retaining one's own income was analogous to abusing alcohol.

From 1935 to 1942, ideological change also proceeded apace, shaped in part by current generations' institutional experiences of the accepted role of government. People increasingly grew accustomed to a more expansive role of government through income taxation, social security, and other governmental programs, despite the political transaction-cost manipulation that spawned those programs. Set against ideologies of earlier times when even a commissioner of Internal Revenue had regarded the income tax as "inquisitorial" and "obnoxious to the genius of our people," the 1943 hearings provided a foretaste of changes already in process. To be sure, many still expressed principled opposition to the income tax. But one can find little in earlier hearings that compares with the following 1943 exchange (U.S. House Hearings 1943: 506) between Representative McLean and ex-Treasury official Elisha Friedman:

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

> Mr. McLean: Do you think there is anything inherently wrong in going too far in compulsory deductions from wages?
>
> Mr. Friedman: I can only come back to this, we have got to do it gradually.
>
> Mr. McLean: Whether you do it gradually or rapidly, I am asking you whether there is anything inherently wrong in taking money out of a fellow's pay envelope without giving him the right to say you are privileged to do it.
>
> Mr. Friedman: Is it wrong for a democratic form of government to do anything? You are the people's elected Representatives. When you decide to do something, it means the people have decided it. *What do you mean, wrong?*

This terse question—and the volumes it spoke about emergent ideological change in the United States—was a harbinger of things to come. Reflecting on the implications of Elisha Friedman's remarks for the American people, McLean (U.S. House Hearings 1943: 506) remarked, "You are trying to take their independence from them."

The pattern of these early events is consistent with the theory under discussion. On one level, the 1943 Current Tax Payment Act was an outgrowth of its institutional predecessors. The political transaction-cost augmentation accompanying adoption of the 16th amendment, tax propaganda during World War I, and implementation of the Social Security Act and Victory tax changed the institutional status quo in ways that reduced resistance to subsequent authority-expanding programs. These early changes increased the government's authority and, with it, its ability to manipulate political transaction costs to further favored policy outcomes such as the 1943 act. Through mechanisms of transaction-cost augmentation, political resistance to these and later measures became increasingly an endogenous product of government officials' institution-modifying choices. Accommodative ideological change ensued.

Later Attempts to Expand Withholding

Similar mechanisms were again evident following 1943's establishment of broad-based income tax withholding. Information reporting on interest and dividend income was established in 1962 (Doernberg 1982). In the late 1970s and early 1980s, efforts were made to expand the withholding system to cover independent contractors' incomes as well as interest and dividend income.

In several dimensions, government officials' later testimony reflected reduced perceived need to conceal their objectives. They wanted tax compliance and said so. Treasury Secretary G. William Miller testified regarding President Carter's proposal to withhold taxes

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

on interest and dividends (U.S. House Hearings 1980: 5) that "the primary purpose of this particular proposal is to improve compliance and to do so on a basis that is practical and economical." Donald C. Lubick, Assistant Treasury Secretary for tax policy, was blunt about it, stating (U.S. House Hearings 1979: 79; U.S. Senate Hearing 1979: 88) that, while "withholding results in high rates of compliance," without withholding "approximately 47 percent of all workers who are treated by payors as independent contractors do not report any of their compensation." IRS Commissioner Jerome Kurtz (U.S. House Hearings 1979: 97) reported that the data "exhibit the basic trend that reporting compliance is highest for income subject to withholding (wages and salaries), somewhat less for income subject to information reporting, and least for income that is generally subject to neither."

Although Treasury officials occasionally reiterated their intention to benefit the taxpayer, such rhetoric was not at the core of the government's argument as it had been in 1943. Compliance was now avowedly the central issue. Brief allusions to taxpayer benefits typically were followed by discussions of compliance. For instance, after proclaiming à la 1943 that "We are proposing no increase in tax, no new tax, no change in taxes . . . merely a change in the method by which individuals will pay the taxes they already owe on their interest and dividends," Treasury Secretary Miller immediately returned to the compliance issue, stating (U.S. House Hearings 1980: 5–6):

> In that area [wages and salaries], where we do have withholding, there is only about a 2- or 3-percent rate of underreporting, while the studies indicate that for interest and dividends from 9 to 16 percent of the taxable income is not reported. This means at least 300 percent greater noncompliance in the [*sic*] areas than in the case of wages and salaries.

Perhaps partly because interest and dividend income was targeted, the present-value issue was more clearly understood by Congress this time around. Time had worked to dispel some of the apparent complexities of the issue. As members of Congress probed, Treasury officials sometimes did not deny that taking people's money sooner hurt them, and instead tried to argue that not much money was at stake for the individual taxpayer. Treasury Secretary Miller, questioned about the withholding proposal's tendency to discourage savings, stated (U.S. House Hearings 1980: 12) that the taxpayer "would *only* lose interest on the amount of the tax that would not have been paid as early in the year if there were no withholding."

His contention that this loss to the taxpayer was negligible again revealed government officials' inclination to misrepresent policy-relevant facts to the public. Miller's strategy was to describe the lost

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

interest income as a small percentage of the overall asset value, thereby obscuring the magnitude of the taxpayer's actual dollar loss. He stated (U.S. House Hearings 1980: 12–13):

> Since the withheld tax on interest paid on a typical savings account averages less than one percent of asset value over the course of the year, at worst the 'loss' of interest on the withheld tax would be less than one-tenth of one percent of asset value.

Supposing an 8 percent annual return on a $1 million investment, if the withholding rate were 10 percent of earnings, $8,000 would be withheld from the taxpayer's $80,000 annual earnings on this investment. Secretary Miller's views notwithstanding, an individual's foregone interest on such withholding would not be negligible; in the aggregate such foregone earnings would be substantial. Astonishingly, in a subsequent response to Representative Richard M. Duncan's (D., Missouri) written questions, Secretary Miller and IRS Commissioner Kurtz (U.S. House Hearings 1980: 34) stated for the record that "For taxpayers who now report the full amount of their taxable interest and dividend income, *there will be no effective change in either their tax liability or the rate of return on their savings.*"

Further attesting to expansion of the transaction-cost-increasing power of government as its authority grows, the 1979 hearing record reproduced a significant excerpt from the Internal Revenue Manual. Under the heading of "Attitude and Conduct of Taxpayer," IRS agents were instructed as follows:

> The file may also contain information received through other channels, such as informant's communications, newspaper items, reports from financial institutions regarding unusual currency transactions, 1099's, revenue agent's information reports, reports from other government or enforcement agencies. . . . Information of this type is of a highly confidential nature. *The agent is cautioned not to reveal to the taxpayer, his/her representative, or any other unauthorized person, that he/she has such other information.* As a precaution it is advisable for the agent not to bring these documents with him/her to any meeting with the taxpayer or his/her representative (U.S. House Hearings 1979: 414).

Such intentional concealment of IRS information and behavior from the taxpayer was coupled with government officials' oft-stated desire to avoid the appearance, but not the reality, of harassment of taxpayers. Treasury Secretary Miller and Internal Revenue Commissioner Kurtz (U.S. House Hearings 1980: 34–35) averred that alternatives to withholding "would require millions of telephone calls, letters and visits, many involving small amounts of tax" which could "easily be regarded as harassment of small taxpayers" and could "generate massive tax-

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

payer resentment and jeopardize our system of voluntary compliance." In subsequent hearings the American Bankers Association asked (U.S. House Budget Comm. Hearing 1982: 100), "Are they not in effect asking the financial industry to do their harassing for them?" For government officials, that was the point. So using the financial industry (as employers had been used since 1943) not only shifted collection costs but also deflected political blame, further raising transaction costs to private citizens of reacting politically to additional federal encroachments upon their earnings.

Contrary to the reality of mandatory third-party extraction of income from wage earners, officials continued to portray the U.S. tax system as grounded in "voluntary" compliance. Commissioner of Internal Revenue Roscoe L. Egger Jr., testified in congressional hearings that "approximately 80 percent of taxes owed are reported and paid *voluntarily* without any IRS enforcement effort at all." Only in his written remarks did he allow that mandatory withholding underlies this "voluntary" behavior, stating (U.S. House Ways & Means Comm. Hearing 1982: 5, 10–11) that "this voluntary compliance results largely from a very workable system of tax administration rules based on withholding and information reporting." As Professor Charles Davenport testified (U.S. House Ways & Means Comm. Hearing 1982: 273), "Our system is said to be one of voluntary compliance, but for some time we have known that compliance is the highest where voluntarism is the least relied upon."

Finally, throughout the 1979, 1980, and 1982 hearings there was clear recognition by government officials of the incremental nature of institutional change. Sheldon S. Cohen, former Commissioner of Internal Revenue, noted that the American public had "gotten used to withholding." Calling withholding the "backbone of the system," he stated (U.S. House Budget Comm. Hearing 1982: 124), "It is not new. It is not a new tax; nobody can complain that it is a new tax." Officeholders viewed incrementalism as instrumental in achieving their long-run aims.[16] Reflecting on the history of withholding—first social security, then the Victory tax, then the 1943 act—Representative Joseph L. Fisher (D., Va.) commented (U.S. House Hearings 1980: 217), "Maybe the moral is to bring this one in gradually."

Gradualism indeed has typified ongoing legislative efforts. Thwarted in the quest for mandatory withholding of interest and dividend income

[16]Regarding a related matter, Commissioner of Internal Revenue Roscoe L. Egger Jr. explained (U.S. House Ways & Means Comm. Hearing 1982: 19–20) that the IRS advocated a mandatory withholding tax on pensions, but that the IRS was promoting information reporting and, initially, voluntary withholding of taxes on pensions as a first step.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

in 1979 and 1980, government officials waited to make the proposal again—in a "crisis" perhaps, or when they could produce a more appealing rationale for the measure, or when the proposal had stronger presidential or party backing.

In 1982 those conditions coalesced. The purported crisis was the budget deficit; the appealing rationale was the "fairness" of using withholding to make tax evaders pay rather than raising taxes on law-abiding citizens; strong executive backing came from President Ronald Reagan. The vehicle was section 301 of the Tax Equity and Fiscal Responsibility Act of 1982 [TEFRA] (P.L. 97-248, 96 Stat. 324, 3 September 1982) wherein Congress authorized 10 percent withholding on interest and dividends with certain exemptions for poor and elderly individuals. Its acknowledged purpose was taxpayer compliance.[17] Representative Daniel D. Rostenkowski (D., Ill.), Chairman of the House Committee on Ways and Means, argued that "collecting taxes from people and from businesses who are now evading taxes is obviously the *fairest* way to produce additional revenue." The ranking minority member of the same committee, Representative Barber B. Conable, Jr. (R., N.Y.), said that President Reagan had "laid his prestige on the line for this measure" (*U.S. Cong. Rec.-House*, 19 August 1982: 6555, 6630).

Transaction-cost augmentation was central to approval of this withholding measure. It was packaged with an omnibus tax bill, increasing the transaction costs to legislators and voters of resisting the provision. Despite repeated congressional efforts to allow a separate vote on the issue, the House of Representatives did not vote separately on the withholding section of TEFRA.[18] Moreover, contrary to article 1, section 7 of the Constitution, the bill's substantive provisions (including the withholding measure) did not originate in the House of Representatives at all but rather in the Senate Finance Committee. The

[17]The Senate Finance Committee, reporting Internal Revenue Service estimates "that 15 percent of dividend income and 11 percent of interest income is not reported by taxpayers" while "99 percent of wage income is reported by taxpayers," concluded that "Withholding improves voluntary compliance" (U.S. Senate Report 1982: 228).

[18]Representative Norman E. D'Amours (D., N.H.) noted (*U.S. Cong. Rec.-House*, 19 August 1982: 22147) that "One hundred and fifty Members of this body have signed a letter to the Committee on Rules asking for a separate vote. An overwhelming number of our constituents . . . oppose withholding of interest and dividend taxes."

Representative John E. Porter (R., Ill.) remarked the following year (*U.S. Cong. Rec.-House*, 17 May 1983: 12493) that withholding "was just one of 96 parts of the last year's tax package, and, unfortunately, we in the House never had the opportunity to vote on the proposal separately. If it had been a free-standing proposal and not part of omnibus legislation, Congress most likely would have overwhelmingly defeated it, just as we did in 1980."

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Senate bill was tacked on to a minor House bill, and the package was sent to conference without House hearings or debate on the bill. The conference bill then returned to the House under a closed rule which precluded amendment or separate voting on its individual provisions. Through procedures described on the House floor as having "abrogated our constitutional responsibilities," the omnibus budget bill strategy permitted the withholding provision to become law despite a House vote two years earlier rejecting 404-4 the study of withholding on dividends and interest (*U.S. Cong. Rec.-House*, 19 August 1982: 22219). *Ex ante*, many in Congress considered tax enforcement through withholding to be less painful than new taxes as a way to narrow the budget deficit.

However, the new withholding provision was not less painful in practice. Public opposition was profound. By August 5, 1983, one month after withholding was to have taken effect under TEFRA, the Interest and Dividend Tax Compliance Act of 1983 (P.L. 98-67, 97 Stat. 369, 5 August 1983) *repealed* TEFRA's provision for withholding on interest and dividends. In its place Congress authorized expanded information reporting coupled with "backup withholding" of 20 percent in specific circumstances involving taxpayer noncompliance.

The repeal of the TEFRA withholding provision reflects the internal dynamics of determinants of transaction-cost augmentation. While variables discussed above impelled congressmen toward use of transaction-cost-increasing strategies to pass TEFRA, many misjudged the intensity of constituent hostility to the withholding provision. But perhaps the measure served its political purpose nonetheless. Since tax compliance provisions accounted for 21 percent of the additional revenue claimed to be generated by TEFRA, the evanescent authorization of interest and dividend withholding allowed Congress the political benefit without the political cost of claimed budget deficit reduction (U.S. House Report No. 98-120, 1983: 2).

Finally, the 1983 repeal of withholding on interest and dividends provides suggestive evidence regarding the long-term influence of transaction-cost augmentation and the path dependency of institutional change. Absent 40 years' experience of withholding, public opinion in 1983 might have opposed wage and salary withholding with equal intensity.[19]

[19]In 1982 Representative Daniel D. Rostenkowski (D., Ill.) remarked (*U.S. Cong. Rec.-House*, 19 August 1982: 6555), "The debate now in progress on interest and dividend withholding occurred 40 years ago on wage withholding. Now wage withholding is a popular system, and perceived as a relatively painless way to accurately pay taxes. I predict that in a few years, we will be able to say the same of interest and dividend withholding."

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Entrenchment of the Machinery of Government

> It is now a minimal problem to maintain a withholding system on salaries and wages, which is absolutely at the heart of our self-assessment technique of paying taxes.
>
> —G. William Miller, Secretary of the Treasury
> (U.S. House Hearings 1980: 23–24)

The question is, can such long-established and carefully contrived practices be reversed—and, more relevantly, are they likely to be reversed? In a representative democracy, the answer to the first question is unequivocally affirmative. Nonetheless, if the theory and evidence presented here are substantially accurate, the answer to the second question is unequivocally negative. The essence of transaction-cost augmentation is guided deflection of certain types of collective political action in order to facilitate establishment and continuation of policies often initially inconsistent with people's preferences. We have seen that the long-run result of such institutional change is authority-legitimating ideological change that renders policy reversal increasingly unlikely.

One sobering result of the present study is its evidence of key public officials' extensive awareness of the dynamics of transaction-cost manipulation. As their published statements made clear, many officeholders deliberately sought the transaction-cost-increasing results of their income tax withholding policies. Along with the public, other officeholders—such as those in 1943 who did not understand the present-value issue—were targets of those transaction-cost-manipulating strategies.

The history of U.S. income tax withholding documented here is consistent with the theory discussed earlier in this article. We have seen that, on many levels, income tax withholding increases transaction costs to the public of understanding the magnitude of the income tax and of opposing it politically. Government officials always have regarded withholding as a seemingly "painless alternative" (U.S. House Hearings 1980: 35). Lacking an understanding of the concept of present value, many taxpayers do not perceive that withholding causes the real burden of their tax liability to be greater. Indeed, the common practice of overwithholding associates the payment of taxes with an apparent financial benefit rather than cost, distorting taxpayers' assessments of the actual costs and benefits of government activity. Consistent with a transaction-cost-manipulation model, the expected return of such overpayments makes people feel "happier" about sending in their tax returns on April 15. The very mechanism of withholding deflects blame from the government by requiring employers to initiate

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

and bear the cost of the forcible extraction of people's income. Piecemeal collection each payday from income the taxpayer never sees obscures the magnitude of the annual tax. And, because it is a forcible extraction, it raises the transaction costs to the public of expressing political resistance to taxes by not paying them.

In examining the process of passing the Current Tax Payment Act of 1943, we have seen that key variables identified by the theory as determinants of transaction-cost augmentation played their expected roles. The complexity of the withholding measure, the appealing rationale provided by wartime revenue needs, executive support, constituent support induced by purported tax forgiveness, ideologies shaped by the already expanded role of government, and lack of media publicity regarding the measure's transaction-cost-increasing features all contributed as expected to adoption of the Act. Consistent with the theory, the history of income tax withholding showed institutional experiences born of transaction-cost manipulation in turn influencing predominant public ideologies, thereby building long-run support for expanded government authority buttressed by additional transaction-cost-increasing institutions.

After 50 years of comprehensive withholding at the source of American workers' salaries, people are used to wage withholding; most no longer question it. The relevant institutional machinery is entrenched, both through its administrative apparatus and through its acceptance in the minds of most taxpayers. Some resistance does remain. Representative Bill Gradison (R., Ohio), for instance, stated (U.S. House Hearings 1980: 46) that "one of the greatest steps we can take toward holding down expenditures and making people aware of the cost to Government would be to reexamine our assumption that wages must be withheld upon." More recently (*Wall Street Journal* 1994), in conjunction with his proposal to replace the existing income tax with a flat tax, Representative Dick Armey (R., Texas) recommended elimination of withholding, calling it a "crucial, deceptive device" that has allowed government "to raise taxes to their current level without igniting a rebellion." But such voices are few. As ideologies accommodate altered institutional reality, as citizens' views about what ought to be come more nearly to reflect what is, government manipulation of political transaction costs provides one key part of the explanation of how such politico-economic change has occurred.

References

Armey, D. (1994) "Review Merits of Flat Tax." *Wall Street Journal*, 16 June: A18.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Baack, B.D., and Ray, E.J. (1985a) "Special Interests and the Adoption of the Income Tax in the United States." *The Journal of Economic History* 45(3): 607–25.

Baack, B.D., and Ray, E.J. (1985b) "The Political Economy of the Origin and Development of the Federal Income Tax." In Higgs, R. (ed.) *Emergence of the Modern Political Economy, Research in Economic History*, Supp. 4, 121–38. Greenwich, Conn.: JAI Press.

Bopeley, J.L. (1943) "Pay-As-You-Go, Civil War Style." *Taxes* 21(7): 376–77, 408.

Brinkley, D. (1988) *Washington Goes to War.* New York: Alfred A. Knopf.

Buenker, J.D. (1981) "The Ratification of the Federal Income Tax Amendment." *Cato Journal* 1(1): 183–223.

Cantril, H. (1951) *Public Opinion 1935–1946.* Princeton, N.J.: Princeton University Press.

Doernberg, R.L. (1982) "The Case Against Withholding." *Texas Law Review* 61(4): 595–653.

Ekirch, A.A. Jr. (1981) "The Sixteenth Amendment: The Historical Background." *Cato Journal* 1(1): 161–82.

Forsythe, D.W. (1977) *Taxation and Political Change in the Young Nation, 1781–1833.* New York: Columbia University Press.

Gallup, G.H. (1972) *The Gallup Poll: Public Opinion 1935–1971*, Vol. 1. New York: Random House.

Higgs, R. (1985) "Crisis, Bigger Government, and Ideological Change: Two Hypotheses on the Ratchet Phenomenon." *Explorations in Economic History* 22: 1–28.

Higgs, R. (1987) *Crisis and Leviathan: Critical Episodes in the Growth of American Government.* New York: Oxford University Press.

Higgs, R. (1988) "Can the Constitution Protect Private Rights during National Emergencies?" In Gwartney, J.D., and Wagner, R.E. (eds.) *Public Choice and Constitutional Economics,* 369–86. Greenwich, Conn.: JAI Press.

Jones, C.C. (1989) "Class Tax to Mass Tax: The Role of Propaganda in the Expansion of the Income Tax during World War II." *Buffalo Law Review* 37(3): 685–737.

Kuran, T. (1987) "Preference Falsification, Policy Continuity and Collective Conservatism." *Economic Journal* 97: 642–65.

Kuran, T. (1991) "Cognitive Limitations and Preference Evolution." *Journal of Institutional and Theoretical Economics* 147: 241–73.

Kuran, T. (1993) "The Unthinkable and the Unthought." *Rationality and Society* 5(4): 473–505.

Leff, M.H. (1983) "Taxing the 'Forgotten Man': The Politics of Social Security Finance in the New Deal." *The Journal of American History* 70(2): 359–81.

Leff, M.H. (1984) *The Limits of Symbolic Reform: The New Deal and Taxation, 1933–1939.* Cambridge: Cambridge University Press.

Lent, G.E. (1942) "Collection of the Personal Income Tax at the Source." *Journal of Political Economy* 50(5): 719–37.

Paul, R.E. (1943) "The Impact of Taxation on Consumer Spending." *Taxes* 21(2): 325–28, 351.

Twight, C. (1983) *Government Manipulation of Constitutional-Level Transaction Costs: An Economic Theory and Its Application to Off-Budget Expendi-*

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ture through the Federal Financing Bank. Doctoral dissertation. Seattle: University of Washington.
Twight, C. (1988) "Government Manipulation of Constitutional-Level Transaction Costs: A General Theory of Transaction-Cost Augmentation and the Growth of Government." *Public Choice* 56: 131–52.
Twight, C. (1989) "Institutional Underpinnings of Parochialism: The Case of Military Base Closures." *Cato Journal* 9(1): 73–105.
Twight, C. (1991) "From Claiming Credit to Avoiding Blame: The Evolution of Congressional Strategy for Asbestos Management." *Journal of Public Policy* 11(2): 153–86.
Twight, C. (1992) "Constitutional Renegotiation: Impediments to Consensual Revision." *Constitutional Political Economy* 3: 89–112.
Twight, C. (1993) "Channeling Ideological Change: The Political Economy of Dependence on Government." *Kyklos* 46(4): 497–527.
Twight, C. (1994) "Political Transaction-Cost Manipulation: An Integrating Theory." *Journal of Theoretical Politics* 6(2): 189–216.
U.S. Congress (1943) *Congressional Record*, Vol. 89.
U.S. Congress (1982) *Congressional Record*, Vol. 128.
U.S. Congress (1983) *Congressional Record*, Vol. 129.
U.S. House of Representatives (23 January 1871) Comm. on Ways and Means, "Income Tax: Letter from the Commissioner of Internal Revenue." House Mis. Doc. No. 51, 41–3.
U.S. House of Representatives (April–May 1941) Comm. on Ways and Means, *Revenue Revision of 1941.* Hearings, Vol. 1, 77–1.
U.S. House of Representatives (March–April 1942) Comm. on Ways and Means, *Revenue Revision of 1942.* Hearings [on H.R. 7378], Revised, Vols. 1–2, 77–2.
U.S. House of Representatives (February 1943) Comm. on Ways and Means, *Individual Income Tax.* Hearings [On a Proposal to Place Income Tax of Individuals on a Pay-As-You-Go Basis], 78–1.
U.S. House of Representatives (28 May 1943) *Current Tax Payment Act of 1943.* Conference Report [To accompany H.R. 2570], House Report No. 510, 78–1.
U.S. House of Representatives (October 1979) Comm. on Ways and Means, *Underground Economy.* Hearings, Serial 96–70, 96–1.
U.S. House of Representatives (30 April–1 May 1980) Comm. on Ways and Means, *President's Proposal for Withholding on Interest and Dividends.* Hearings, Serial 96–92, 96–2.
U.S. House of Representatives (19 March 1982) Comm. on the Budget, *User Fees and Withholding Taxes on Interest and Dividends.* Hearing, 97–2.
U.S. House of Representatives (18 May 1982) Comm. on Ways and Means, *Tax Compliance Act of 1982 and Related Legislation.* Hearing [On H.R. 6300], Serial 97–63, 97–2.
U.S. House of Representatives (13 May 1983) Comm. on Ways and Means. House Report No. 98-120 [To accompany H.R. 2973].
U.S. Senate (19 July 1894) *Estimate of the Probable or Possible Revenue under the Proposed Income Tax.* Senate Mis. Doc. No. 232, 53–1.
U.S. Senate (16 June 1909) "Tax on Net Income of Corporations." Message from the President of the United States, Senate Doc. No. 98, 61–1.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

U.S. Senate (16 February 1910) "Message from the Governor of New Jersey Transmitting to the Legislature the Proposed Sixteenth Amendment to the Constitution of the United States Relative to the Income Tax, February 7, 1910." Senate Doc. No. 365, 61–2.

U.S. Senate (14 December 1910) *Shall the Income-Tax Amendment Be Ratified?* Senate Doc. No. 705, 61–3.

U.S. Senate (July–August 1942) Comm. on Finance, *Revenue Act of 1942*. Hearings [On H.R. 7378], Vol. 1, 77–2.

U.S. Senate (19 August 1942) Subcomm. of the Comm. on Finance, *Withholding Tax*. Hearing [On Data Relative to Withholding Provisions of the 1942 Revenue Act], 77–2.

U.S. Senate (6–7 May 1943) Comm. on Finance, *Current Tax Payments Act of 1943*. Hearings [On H.R. 2570], Revised, 78–1.

U.S. Senate (10 May 1943) *Current Tax Payment Act of 1943*. Senate Report No. 221 [To accompany H.R. 2570], Parts 1–2, 78–1.

U.S. Senate (1946) Comm. on Finance, *Legislative History of the Current Tax Payment Act of 1943*. Comm. Print, 79–2.

U.S. Senate (17 September 1979) Comm. on Finance, *Miscellaneous Tax Bills II*. Hearing, 96–1.

U.S. Senate (12 July 1982) Comm. on Finance, *Tax Equity and Fiscal Responsibility Act of 1982*. Senate Report No. 94–494 [To accompany H.R. 4961].

U.S. Treasury Department (1916) *Annual Report of the Secretary of the Treasury on the State of Finances for the Fiscal Year Ended June 30, 1915*. Washington, D.C.: U.S. Government Printing Office.

U.S. Treasury Department (1917) *Annual Report of the Secretary of the Treasury on the State of Finances for the Fiscal Year Ended June 30, 1916*. Washington, D.C.: U.S. Government Printing Office.

U.S. Treasury Department (1918) *Annual Report of the Secretary of the Treasury on the State of Finances for the Fiscal Year Ended June 30, 1917*. Washington, D.C.: U.S. Government Printing Office.

U.S. Treasury Department (1919) *Annual Report of the Secretary of the Treasury on the State of Finances for the Fiscal Year Ended June 30, 1918*. Washington, D.C.: U.S. Government Printing Office.

U.S. Treasury Department (1944) *Annual Report of the Secretary of the Treasury on the State of Finances for the Fiscal Year Ended June 30, 1943*. Washington, D.C.: U.S. Government Printing Office.

Witte, J.F. (1985) *The Politics and Development of the Federal Income Tax*. Madison: University of Wisconsin Press.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The Entrenching of Incumbency: Reelections in the U.S. House of Representatives, 1790–1994

Stephen C. Erickson

The "Intuitive Impulse"

On November 8, 1822 an article on term limits, or "Rotation in Office" as they were called, appeared in the *Richmond Enquirer*. The author argued that long tenure in public office leads to abuse of power. He declared that the principle of rotation of public offices was "long cherished" in America and was impressed upon the public mind "by a kind of intuitive impulse, unassailable to argument or authority."

The overwhelming popularity of the current term limits movement suggests that this sentiment remains alive and well in modern American political culture, despite very high incumbent reelection rates. As a general principle, Americans continue to believe in a centuries-old democratic tradition that power held over extended periods of time corrupts representative institutions. And today, many Americans seem convinced that their political institutions are thoroughly corrupt.

Nevertheless, of those who choose to run for reelection in the U.S. House of Representatives, over 90 percent are assured of victory, a fact it is easy to lose sight of given the historic Republican takeover of both houses of the U.S. Congress in the 1994 election. The Republican "revolution" was based on their ability to toss a small percentage of Democratic incumbents out of office, and to win open seats. In fact, the power of incumbency masked the true scale of the popular sentiment in favor of Republicans. The advantages of incumbency not only make elections unfair to challengers, but also minimize the impact of popular partisan swings on Congress.

This paper traces the path of reelection rates and congressional turnover historically, through the fertile valleys of citizen legislatures

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

The author is a former Resident Scholar at U.S. Term Limits and a doctoral candidate in Early American History at the College of William and Mary.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

to the mountains of immovable incumbency. The focus is on the U.S. House of Representatives, since Senators were subject to indirect election up until after the turn of the 20th century. The paper is an attempt to provide historical context for the current state of American democracy, and in so doing explain the peaks and valleys of House reelection rates, as illustrated in Figure 1. Indeed, Figure 1 is a kind of topographical map of incumbency to accompany the following description of the reelection landscape since the nation's founding.

The history of congressional turnover readily breaks into four parts: (1) the Republican Era of Social Deference and Public Service, 1787–1828; (2) the Era of Jacksonian Democracy, 1829–95; (3) the Era of Safe Seats, 1896–1957; and (4) the Era of Entrenched Incumbency 1958–present. Each of the four periods represents a distinct set of circumstances that shaped elections and reelections for the U.S. House of Representatives.

The Republican Era of Social Deference and Public Service

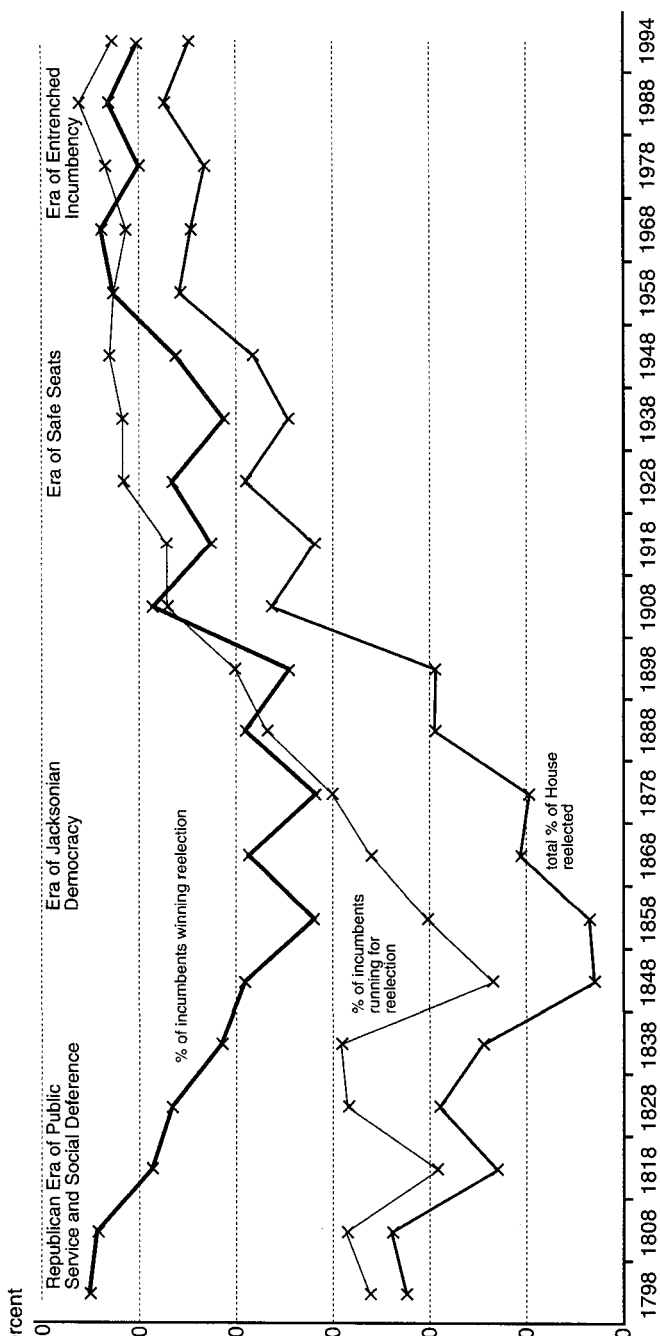
James Madison was immersed in political theory during the winter of 1786–87. He read voraciously prior to the Constitutional Convention that met in Philadelphia the following spring. Although the convention was billed as a means to improve the existing Articles of Confederation, Madison and some of his allies had far more ambitious plans for the gathering. They were going to construct a new political system from the ground up, that would usher in "a new order of the ages" (Novus Ordo Seclorum). Madison and his allies understood that the success of the great American republican experiment rested on their ability to properly structure the new government. His political consultants included the great philosophers of antiquity, generations of English republican theorists, and the leading figures of the European Enlightenment. Among the subjects Madison's reading encouraged him to consider, when he contemplated the proper design for a national legislature, was "rotation in offices."

The principle that new representatives should regularly be circulated through the political system was an integral part of republican political theory. Aristotle (*Politics*, IV, Sec. 2: 2091) wrote approvingly of how office rotation was practiced in ancient Greece, where citizens were retired from public office to live under the laws they had made, and replaced by a new group of citizens. Livy recorded that offices were also rotated in the Roman Republic. Sixteenth-century Italian humanist writers advocated republican political structures, including mandatory rotation of offices, that encouraged rulers to exercise virtue

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

FIGURE 1

U.S. HOUSE REELECTION RATES


Percent
100
90
80
70
60
50
40
Republican Era of Public Service and Social Deference
Era of Jacksonian Democracy
Era of Safe Seats
Era of Entrenched Incumbency
% of incumbents winning reelection
% of incumbents running for reelection
total % of House reelected
1798 1808 1818 1828 1838 1848 1858 1868 1878 1888 1898 1908 1918 1928 1938 1948 1958 1968 1978 1988 1994

NOTE: Reelection rates are based on five election cycle averages, except for the mose recent period: 1990–94.
SOURCE: Congressional Research Service.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

over self-interests. From the Italian humanists the Englishman James Harrington adopted many of his political ideas. Like James Madison much later, Harrington aspired to design and build a republic. The result was his *The Republic of Oceana* ([1656] 1977), a constitutional blueprint that Harrington tried, unsuccessfully, to sell to Oliver Cromwell during the English Civil War. For Harrington, rotation of offices was absolutely essential to the well-being of a republic. A legislator returning to the body of the people was analogous to a blood cell returning to the lungs to be enriched with fresh oxygen (Pocock 1975: 393–95). Generations of English republican thinkers adopted Harrington's ideas on rotation (Robbins 1968). John Trenchard and Thomas Gordon, who were especially influential in America, wrote on the subject in *Cato's Letters* ([1755] 1995: Nos. 43, 61; Bailyn 1967: 35–36) and warned that long tenure in office frequently led to aristocracy.

Madison was familiar with all of these political philosophers, and no doubt appreciated their arguments for mandatory rotation of offices. He was also well acquainted with the various laws of the American states. Not only had there been rotation under the Articles of Confederation, but many states incorporated term limits in their early constitutions. Virginia, Pennsylvania, Maryland, and Delaware all rotated their governors out of office. Virginia, New York, and Delaware limited the number of terms that Senators could serve. Most radically of all, Pennsylvania limited the terms of the members of its unicameral legislature. Other states regularly rotated sheriffs and coroners (Adams 1980: 308–9). In history, in theory, and in practice, Madison and his fellow constitutional framers had many models from which to borrow concepts for republican government, and many of these included term limits.

Behind the scenes Madison prepared the Virginia delegation for the Constitutional Convention to be held at Philadelphia in 1787. In letters to George Washington (16 April 1787) and Edmund Randolph (8 April 1787), Madison outlined what would become known as the Virginia Plan, the working draft of the Constitution itself. In the Virginia Plan, Madison included radical term limits for the House of Representatives, consisting of mandatory retirement after a single two-year term in office (Madison's "Notes on the Federal Convention" in his *Papers*).

Madison made a unique contribution in defense of term limits. He prophetically warned of the danger that factions, or special interests, presented to a republican form of government. He worried that legislators, absorbed by their desire for reelection, might serve their own narrow self-interest at the expense of the overall national interest. By

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

mandating periodic retirement for legislators, and therefore making reelection impossible at times, Madison hoped to minimize the problem of special interest politics (Erickson 1993).

Madison may have worked with fellow Virginian George Mason when drafting the Virginia Plan. Mason was an ardent believer in rotation of offices. He had included term limits in the widely imitated Virginia Declaration of Rights, which suggests that like other specifically enumerated rights, mandatory term limits were considered an essential guarantee against tyrannical government. Mason was disappointed by the exclusion of term limits from the Constitution, along with other provisions that would later be incorporated as the Bill of Rights. Likewise, upon reading the Constitution for the first time, Thomas Jefferson declared his hope that some of its shortcomings might be remedied. "One thing I dislike, and greatly dislike," Jefferson wrote to Madison (20 December 1787), "is the abandonment in every instance of the necessity of rotation in office." The absence of any assurance that there would be rotation of office was also a sticking point for many Americans who would become known as Anti-Federalists.

The provision for term limits was dropped, without discussion, from the Virginia Plan. Madison would again briefly return to the subject during the course of the debate over the Constitution at Philadelphia. Yet, it appears that among the framers, Madison's and Mason's view on the necessity of mandated rotation was decidedly in the minority. Term limits were also not included in the Bill of Rights, where Mason might have incorporated them. It would be a mistake, however, to assume that the framers favored professional politicians. The historical context in which the framers designed their republic helps explain the exclusion of term limits.

The 13 states emerged from the Revolution victorious, but hobbled with enormous financial difficulties. They faced a mountain of public and private debt, as well as a severe economic depression. Specie drained out of America into the coffers of large merchants in England, individual states were insolvent, foreign loans needed to be paid off, and internal trade barriers between states discouraged commerce. Western Massachusetts erupted in outright rebellion. The government created by the Articles of Confederation was hopelessly too weak to adequately address the young republic's formidable problems.

To the gentlemen who wrote the Constitution, politics had also taken a dangerous turn after the Revolution. They saw demagogues appealing to the passions and narrow self-interests of the masses. Men who had formerly been of low social stature in their communities were now exalted to high public office, while representatives of the

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

most elite families shunned public service along with the new brand of relatively democratic politics.

No radical development better illustrated the fears of elite republican culture than the adoption of the Pennsylvania Constitution of 1776, which included a unicameral legislature, radical suffrage, and mandatory rotation in office. The new Pennsylvania Constitution swept the old Quaker elite from power, and replaced them with radical evangelical Scots-Irish. Arguably, Pennsylvania's legislative term limits provision was less radical in practice than it was in spirit. It openly distrusted those elite characters who would normally be elected to serve in government, implying that they would not necessarily retire of their own volition. Thus rotation was forced through the power of law (Wood 1969).

In sum, the authors of the Constitution failed to include a provision for mandatory rotation in office because they perceived that the "best" men were no longer serving in government. New unsophisticated and ignorant politicians had, in many cases, come along to take the place of America's leading patriot families. By creating large congressional districts, the framers hoped that men would win office based not on their political connections or demagoguery, but on sterling public reputations known far and wide. Simply put, they hoped that under a new system, men like themselves would be elected. Given the public's apparent predisposition to throw gentlemen out of office, term limits seemed at least unnecessary, and at most counterproductive.

While in retrospect the framers sound elitist—indeed, they were elitists—their attitude was a product of the time and place in which they lived. The world of 18th-century America, even within most republican circles, was strictly hierarchical. The best families asserted their social positions in every way possible. They lived in the nicest homes, wore elegant clothing, sat in the front pews of churches, and served in public office. Tipping their hats in deference to their community's "natural leaders," common folk generally went along with this hierarchical system. To be sure, significant social mobility existed in America, but only successful men who carried themselves as gentlemen could assume leadership in the community.

The framers hoped to encourage the nation's leading men to serve in government. Naturally, the prominent men were wealthy. Wealth meant financial independence. A gentleman of means, it was believed, would be more likely to serve the needs of his community and was less tempted by the corrupting influences of money and power associated with high public office. Furthermore, it was quite natural that only those with the leisure time to read broadly could govern justly. There was no congressional staff, no think-tanks, and no bureaucracy

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

to guide elected representatives. There were few books around by today's standards, and those that existed were expensive items. Colleges and universities stressed theology. Practical training for running a government had to come on the job, without the help of public policy experts. Given the complexity of the economic and political problems in post-revolutionary America and the diversity of local conditions throughout the 13 states, the new nation seemed to need experienced legislators.

However, the framers did not favor professional politicians, whose motivations could not be disinterested. Looking back on the long struggles of English Whigs against the Stuart monarchies, American republicans understood the danger that sprang from a government class, self-serving, and distant from the people. The generation that made the American revolution, both Federalist and Anti-Federalist alike, scorned those who made their living from government. The ideal republican was embodied in Cincinnatus, a Roman farmer who put down his plow to take up arms in defense of the community. When the crisis ended, he returned to his plow. The motivation driving Cincinnatus was public virtue; not money or power, but a reputation as a virtuous citizen, was the permissible reward under a republican system of government. George Washington's behavior toward public office—his outward reluctance to hold power, climaxing in his refusal to serve a third term as President—exemplified the ideal of Cincinnatus.

The political values of 18th-century republicanism faded slowly and continued to dominate electoral behavior well into the 19th century: A high turnover rate in Congress, with about one-third of its members replaced each election cycle, was precisely the rate James Harrington and his many American political heirs had called for in mandatory rotation laws. Madison did not fight very hard for term limits, probably because he knew that the sentiment at the Philadelphia Convention was against them, and the specter of entrenched incumbency was then only a theoretical menace. In the early years of the republic, Madison's hopeful speculation proved true, that a "few" members of the legislative branch would be "frequently reelected" (*Madison's Notes*, 21 June 1787), but "new members would always form a large Proportion" (*Federalist No. 53*). Concerns that Congress would become a separate, entrenched class seemed unwarranted. It appeared that no term limits law was necessary.

The high turnover rate in the late 18th and early 19th centuries, as Figure 1 shows, was dependent on regular voluntary retirement from public office. Certainly the spirit of Cincinnatus, which played an important role in the political culture, encouraged such behavior.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

However, public service in early America was in many ways a genuine hardship. Living costs and social expectations at the capital outweighed the financial compensation for serving in Congress. Most Congressmen probably left Washington, D.C. less wealthy than when they arrived. Poor transportation, consisting primarily of notoriously bad roads, made trips to the banks of the Potomac an arduous journey. Once in Washington, trips home for business or to visit family were necessarily few and far between. The incentives to remain in the District of Columbia were also underwhelming. Hardly a bustling metropolis, the capital consisted of only a few buildings scattered about in a physical and social wilderness. The activities of the federal government were often mundane, and arguably, the most important decisions were made at the state level throughout much of the 19th century. In fact, members of Congress would often return home to serve in state legislatures (Kernell 1977).

For all of these reasons, Congress experienced a high rate of retirement in the early republic. Turnover had much less to do with defeat at the polls, especially in the years immediately following the establishment of the Constitution. The egalitarianism of the Pennsylvania Constitution was premature. It was replaced by a more conservative document, made in the mold of the Massachusetts and federal Constitutions. Well into the 19th century, leading families continued to send representatives to serve in government, who were dutifully elected and reelected, when they chose to stay in office, by their fellow citizens occupying lower steps on the social ladder. Turnover remained high due to frequent voluntary retirements. Yet the American political system increasingly accommodated itself to more egalitarian political ideas and practices (MacNeil 1963). As the largest-scale democratic meetings ever assembled in America, the state conventions that ratified the Constitution were themselves radical departures in self-government created by the federalists (Wood 1969). Indeed, the new social and political order planted at the time of the Revolution took root slowly, but its growth was unmistakable. To trace the declining rate of those who stood for reelection and won between 1800 and 1850 in the U.S. House, is to mark the decline of a system of social and political deference, and the rise of values commonly associated with Jacksonian democracy.

The Era of Jacksonian Democracy, 1829–95

The language and ideas of the American Revolution created an ideology more politically and socially radical than many of the makers of the new republic intended. The notion that "all men were created

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

equal," for example, could be taken far beyond the basic principle that all men merely had equal rights. Political equality, taken to an extreme, might mean that all men were equally capable of governing as well as being governed.

While Andrew Jackson considered himself a political heir to Jefferson, his attacks on "privilege" and the growing suspicion that no representative could legislate with disinterest made the whole idea of "public virtue" appear almost chimerical. The intensified skepticism over the ability of any officeholder to remain in power and uncorrupted for any significant length of time led to comprehensive applications of the idea of rotation of offices.

To be sure, members of the revolutionary generation, like Madison and John Adams, held no illusions about human nature. What changed was the notion that elite gentlemen were better equipped to legislate for the public good, and were naturally more disinterested. Americans became steadily more convinced that all power corrupted, and that wealthy planters and lawyers were as self-interested and corruptible as anyone else. Elite political leaders could not be counted on to legislate for the common good or represent the common man. The idea that gentlemen allegedly had a monopoly on reason was replaced by a widespread faith in common sense (Wiebe 1984, Wood 1993).

Andrew Jackson (as quoted in Richardson 1896: vol. 2, 448) said as much in his address to the 21st Congress,

> There are perhaps few men who can for any great length of time enjoy office and power without being more or less under the influence of feelings unfavorable to a faithful discharge of their public duties. The duties of all public officers are, or, at least, admit of being made so plain and simple, that men of intelligence may readily qualify themselves for their performance; and I cannot but believe the more is lost by long continuance of men in office, than is generally gained by experience.

Jackson's deep distrust of the holders of government-granted power and privilege led to his well-known war on the Bank of the United States, and attacks on other government monopolies. Jackson's presidency also marked the beginning of the high point of office rotation, "a leading principle of the Republican Creed," as Old Hickory called it.

The number of House members who sought reelection and won after 1800 declined over the course of the next half century as the era of citizen government reached its height. Incumbents who chose to run again were 95 percent successful at the beginning of the century, when old hierarchical and deferential social structures remained in place. But in the election of 1842 the survival rate of incumbents bottomed out at 64 percent. That same year so many incumbents

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

chose not to run for reelection that Congress experienced a 76-percent turnover. By the 1840s, the reelection rate for incumbents attempting to maintain office settled into an average range where 20 to 30 percent regularly went down to defeat. This attrition rate due to defeat continued almost up until 1900. Adding in voluntary retirements, the total turnover rate in Congress averaged more than 50 percent per election between 1840 and 1880.

Many of the founding fathers would have looked askance at the U.S. House of Representatives of the mid-19th century, dominated as it was, not by the nation's leading characters, but by quite ordinary men. The famous French traveler and aristocrat, Alexis de Tocqueville, while admiring much of what he saw in America, was disdainful of the lower house of the United States legislature. "On entering the United States House of Representatives in Washington," sniffed de Tocqueville ([1835] 1965: 103), "one is struck with the vulgar demeanor of that great assembly." He added that "the eye frequently does not discover a man of celebrity within its walls." The body contained "almost all obscure individuals," including "village lawyers, men in trade, or even persons belonging to the lower classes of society." The U.S. House of Representatives in the mid-1800s came as close as it ever would to a true citizen legislature. In the same report, de Tocqueville pointed out that the Senate, by contrast, contained gentlemen who compared favorably with their European counterparts. Madison and Jefferson would have at least appreciated the contrast between the two houses. Madison's goal of a lower house that was immediately reflective of the people was realized in the 19th century.

While the belief in social and political deference that existed during the age of the nation's founding faded, the republican principle of voluntary rotation did not. In fact, the custom became more prevalent as a result of the rise of party politics and the birth of "the spoils system" (McCormick 1986) under Jackson. The spoils system constituted a new form of rotation, where tenure in appointed office became limited by the duration of the authority of the appointee's political allies. When Jackson spoke of the disadvantages of allowing individuals to hold public office for extended periods of time, he meant all public offices, including administrative ones. Since all offices were public property, they were free to be distributed to the party faithful. According to the Jacksonians, such a rotation not only rewarded political allies, but also discouraged the corruption that stems from long tenure in office.

Acceptance of the idea of office rotation became so dominant during the 19th century that a set of unwritten rules evolved. In district political conventions, party leaders took turns holding elected office.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

In 1824 the *New York Statesman* (as quoted in Struble 1979: 659) reported that "perhaps most" of the turnover in the state's congressional delegation was the result of "an arrangement, by which it is stipulated that, after a given time, one aspirant for office is to succeed another."

Often counties took turns sending representatives to Washington, in a geographic rotation. In 1869, the *Christian Examiner* (as quoted in Struble 1979: 660) noted the "constant temptation, in a district made of an aggregation of counties or towns, to pass the office from town to town, or county to another each claiming in its turn the honor of furnishing the member."

One of the most well-known examples of the practice of rotation involved Abraham Lincoln. In describing Lincoln's 7th Illinois Congressional District, his secretaries, John Nicholay and John Hay, reported (as quoted in Struble 1979: 659–60),

> The Sangammon district was one which the Whigs of Illinois had apparently the best prospect of carrying, and it was full of able and ambitious men, who were nominated successively for the only place which gave them the opportunity of playing part in the national theater at Washington.

Lincoln, by agreement, served only one term in Congress. When the subject of running for reelection came up, Lincoln wrote (as quoted in Struble 1979: 660), "to enter myself as a competitor of another, or to authorize anyone so to enter me, is what my word and honor forbid." The man who succeeded Lincoln, Steven Logan, was the same man who had nominated Lincoln two years earlier.

The strong party system of the 19th century contributed to rotation of offices by encouraging compromises between politicians and different geographic regions within congressional districts. A candidate's fate was closely tied to the fate of his political party and the presidential candidate that headed the ticket. Therefore a candidate would not campaign directly for himself, but for his party, on which his own political success depended. Major campaign contributions went through party channels to the candidates. Party control through money and closed nominating conventions helped party bosses to enforce rotation of offices. Rotation itself contributed to party discipline and coherence by allowing various members to play a role on the national stage in the U.S. Congress.

Yet the party system and the rotation of offices that often went along with it underwent dramatic changes at the end of the century. Such changes were ultimately the products of the tumultuous transformations occurring in the American economy and society toward the end of the 19th century.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The Era of Safe Seats, 1896–1957

The transformation of the agrarian republic into the industrial nation-state radically shook American social and political structures, including the electoral system. Industrialization, urbanization, sectionalism, the rise of a new ideology of professionalism, and large-scale immigration, combined to undermine the old practice of rotation of offices and made elections less competitive. Marginal congressional districts—districts rather evenly balanced between Republicans and Democrats—began to disappear, creating uncompetitive seats in both the South and the North. The era of safe seats was caused first, by a polarization of economic interests, and second, by a systematic disenfranchisement of minority groups during the progressive era.

Distinct sectional interests, given bold definition by northern industrialization as well as by southern slavery and its aftermath, dominated American politics both before and after the Civil War. The issue of the tariff in particular cleaved the nation into sections, with the industrial North and Midwest favoring protectionism, and the South and far-West opposed. Agricultural regions also favored a silver standard to ease debt, which was opposed by Eastern financial interests. The Populist political revolt of the 1890s underscored the sectionalist nature of the economic conflict. The Populists drew their strength from agricultural areas in the South and West. The election of 1896, when the Democrats nominated the Populist William Jennings Bryan as their standard-bearer, is often characterized as one of the great realigning elections in U.S. history (Bensal 1984, Kleppner 1987). While the effect of that particular election may be overstated, there is little doubt that the last decade of the 19th century, and first decade of the 20th century, saw the solidification of Republican strength in the North and a Democratic monopoly in the South. Clearly much of this polarization was based upon economic sectional differences.

With depressions setting in like plagues, huge numbers of non-Protestant immigrants pouring into overcrowded cities, labor unrest, urban corruption, and unhealthful living and working conditions, the scale of changes taking place was alarmingly unprecedented. One historian has aptly named the response to this upheaval "a search for order" (Wiebe 1967). Many progressive reformers believed that the chaos of the time required professional management. J.P. Morgan, the great financier, argued that government should be run like a business.

Historian Robert Struble contends that an ideology of professionalization replaced the belief in office rotation and slowed turnover in the U.S. House at the turn of the century. Certainly a waning of the practice of rotation accompanied the trend toward professionalization. However, Americans did not necessarily abandon the ideal of citizen

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

government in favor of a professional class of lawmakers. The disappearance of voluntary rotation had other causes as well.

The radical transformation of political parties at the end of the century, brought about by the progressive belief that old party practices and structures were fundamentally corrupt, was a major reason for the change. The introduction of open primaries and the overall weakening power of party bosses removed policing forces that had helped enforce rotation of offices. Party leaders could no longer limit an officeholder's tenure, because the power to do so increasingly rested with the electorate, and, as we shall see, the electorate was increasingly configured in such a way as to discourage competition at the polls. With the dismemberment of old party structures, a candidate's electoral fate was, in part, cut free from the success or failure of the party as a whole.

Yet the decline in voluntary rotation began before both the rise of the ideology of professional management and well before the beginning of party decline. The growth of cities provides another explanation for the long-term decline of voluntary rotation. The process of intense urbanization began in many areas, especially in the Northeast, before midcentury. After 1910, rural America could no longer marshal a majority in the federal census. Not surprisingly, politically powerful urban centers had little enthusiasm or incentive to share power with sparsely populated agricultural communities. Naturally, as some congressional districts became heavily weighted in favor of urban concentrations, the imbalance discouraged geographic rotation of offices.

Native-born white Protestant rural America responded with alarm to the growth of ethnic urban centers. That agricultural America might lose its place of dominance had profound implications for a nativist culture still infused with a Jeffersonian agrarian ideal. Nativist anxieties about the effects of a growing ethnic urban influence on American democracy prompted the old Protestant classes to entrench themselves, politically, against newcomers who, it was believed, were generally unfamiliar with the habits of democracy. The nativists' first line of defense was to protect their interests through gerrymandering.

Of course, gerrymandering was nothing new. But by the end of the 19th century, state legislatures, when revising political districts, increasingly turned away from drawing lines based on population. Instead they invoked a "federal principle," which justified overrepresentation for rural communities, which in turn tended to be white and Protestant. In defense of the federal principle, Elihu Root, arguing at the New York Constitutional Convention of 1894, pleaded that his state follow the lead of other states with "great cities" within their

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

borders. He advocated putting small agricultural communities on an equal footing with urban centers. "Otherwise," declared Root (as quoted in Dixon 1968: 83–84), "we can never have a truly representative and truly republican government."

In the last decade of the 19th century, the federal principle created congressional districts in various states that were 40 percent larger than the state's smallest district (Dixon 1968: 91). The consequences of such redistricting isolated immigrants, who tended to be Democrats, in burgeoning urban districts. Hence, in the North, the federal principle encouraged proportionally fewer Democratic urban districts and more rural Republican ones. Likewise, the principle was one of several tools employed to disenfranchise blacks in the South, which meant undermining the Republican vote below the Mason-Dixon line. Disparities between the populations of urban and rural districts increased during the course of the 20th century until the Supreme Court intervened in 1964. Following the landmark Supreme Court case *Baker v. Carr*, which found unequal representation in state legislatures unconstitutional, *Wesberry v. Sanders* put an end to the "federal" apportionment system for U.S. House seats. The case involved congressional districts in Georgia, where the state's smallest district contained 272,154 rural inhabitants, compared with the 823,860 inhabitants of Georgia's largest district in Atlanta (Dixon 1968; *Congressional Quarterly* 1985: 694–95).

Redistricting was just one way Jim Crow established safe Democratic seats in the South. Northern efforts to reconstruct the South into a biracial democracy ended in 1877 when federal troops withdrew from the old confederacy. Yet Republicans won seats in the South through 1900. Sometimes as many as five or six Republicans represented North Carolina or Tennessee. However, in 1893, when the Democrats controlled the federal government, it formally abandoned supervision of federal elections in the South. Turning the South's black citizens into scapegoats, Southern populists and "Bourbon Democrats" firmly united under the banner of Jim Crow in 1896. In the years that followed, Southern blacks, who were predominantly Republicans, were systematically disenfranchised through the imposition of poll taxes, property qualifications, and literacy tests. With the exception of East Tennessee, the South was solidly Democratic by 1910, and remained so through the middle half of the 20th century (Woodward 1955, Grantham 1988).

One of the ironies of the period was that as blacks were disenfranchised, women obtained the vote. However, even this expansion of democracy undermined marginal districts by swelling Republican majorities in the North.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Women's suffrage was a long time coming. It grew out of abolitionism, and for most of the movement's 70-year history, based its appeal on the notion that all people were created equal and worthy of full citizenship rights. Yet as historian Aileen Kraditor discovered, women's suffrage only succeeded when, at the end of the century, suffragists began appealing to Protestant nativism. They won over support for their cause by arguing, first, that there were more native-born women than foreign-born men and women combined. Second, they advocated literacy tests, which would further exclude the foreign born from politics and further entrench the literate, native-born Protestant class (Kraditor 1965).

Literacy tests could be defended everywhere as a progressive reform to ensure an educated citizenry. However, in the North, like the South, literacy tests furthered the interests of native-born elites. Nine Northern states imposed such tests. As black Republicans were disenfranchised in the South, literacy tests and women's suffrage undermined Democrats in the North.[1]

The era of safe seats also saw the birth and expansion of a number of practices associated with modern government, including the creation of legislative staff. There is, however, little evidence to suggest that such practices had much of an effect on congressional turnover during the first half of the 20th century. But the practice of voluntary rotation took a near fatal blow. As a result of the destruction of party power and the effects of urbanization, incentives that encouraged voluntary rotation disappeared. Sectionalist economic interests furthered the creation of safe seats, as did progressive era reforms that entrenched white Protestant Republicans in the North and Democrats in the South.

In 1910 dissident Republican members of Congress joined their Democratic peers in a successful revolt against Republican Speaker Joseph Cannon. In so doing they undermined the power of the Speaker of the House and created a congressional system based on seniority. It is noteworthy that Republicans broke with their party in order to create a system that furthered their own individual political interests. Longevity in office, not party loyalty, became the new route to power. Amid the chaos of the period, the new direction in American politics may have appeared uncertain. However, from a historical distance, it

[1]A statistical analysis by David W. Brady and Bernard Grofman (1991), who compare district swing ratios over time, supports the notion that safe seats were multiplied around the turn of the century and that progressive reforms were largely responsible. While Brady and Grofman note the effects of Jim Crow in the creation of safe Southern seats, their data also reveals the proliferation of safe seats in the North after the passage of women's suffrage.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

is clear that structural developments, many of which were still in their infancy, were beginning to systematically erode American democracy.

The Era of Entrenched Incumbency, 1958–Present

While the era of safe seats undermined marginal districts, part of an incumbent's safety arose from a natural, as well as an artificial, distribution of regional interests. Incumbents were safe because more often than not, they represented their constituent's interests, even if those constituencies were subject to some manipulation. Furthermore, in times of crisis, during the Great Depression, for example, more than one-third of the U.S. Congress might fail to be reelected. Yet by 1960, the power of incumbency itself, as distinct from a constituency manipulated to be homogeneous, began to climb. Burgeoning congressional staffs, massive use of the franking privilege, growing incumbent campaign war chests, mass media, and expanding federal power combined to provide incumbent members of Congress with advantages that made them almost immune to defeat.

Ground was broken for incumbent entrenchment early in the century. A few astute observers recognized the potential for abuse in what was then a new congressional system. De Alva Stanwood Alexander, a scholar of Congress writing in 1915, foresaw the dangers of new congressional resources, especially staff. "A member's clerk," declared Alexander ([1915] 1970: 152), "becomes the head of a campaign bureau, with the hope that franked favors may induce grateful constituents to remember the sender on primary and election days." Alexander may have fretted excessively over the impact of a single clerk and the franking privilege. Yet clearly politicians understood from very early on how to use public office to secure reelection. If Alexander's concerns seemed overly anxious in 1915, by 1975 they were prophetic.

Much has been written about the more recent lack of congressional turnover since political scientists like Erikson (1971), Mayhew (1974a, 1974b), and Fiorina (1977) began noticing the trend in the early 1970s.[2] There is general agreement that money from political action committees (PACs), institutional perks, and federal pork have all contributed to the preservation of incumbency. Many scholars have rightly pointed to the increasing role and power of the federal government to explain the transformed nature of congressional priorities. Acting as ombudsmen, members channel resources to their districts, and perform constituent services in order to help secure continual reelections. How-

[2]For a collection on much of the best scholarship on the subject, see *The United States Congress: The Electoral Connection, 1788–1989* (1991). See also the other volumes in the series for a thorough review of periodical literature on the U.S. Congress.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ever, the entrenchment of incumbency appears to have begun at a rather static time in the expansion of federal authority and power. Precisely when and how incumbency began its entrenchment remains a point of some contention, but the broad outlines of the process are clear.

Political scholars have devised one helpful approach to distinguish whether it is the power of incumbency that reelects officeholders, or the safety of their seats. In other words, does a politician continue to get reelected because he or she represents like-minded people, or does the politician use the power of public office to gain advantage at election time? John Alvord and David Brady (1989) address this question by charting two measures. First, they take the margin of an incumbent's victory in his or her last reelection and subtract the margin of victory of the winner of the following open-seat election, provided that the open-seat election is won by a member of the same party. They also calculate the difference between the winner's margin in an open-seat race and his or her first reelection. Since the open-seat margin is almost always smaller, these measures are called the "freshman slump" and the "sophomore surge," respectively. Alvord and Brady average out these differences for each election cycle beginning in 1846. In this way they separate the effect of incumbency from the natural safety of the district. As illustrated in Figure 2, the results are rather dramatic. While there has always been an advantage to being an incumbent—usually it provided two or three extra percentage points at the polls—at about 1958 that advantage took off to levels of seven and eight points, and has recently soared even higher. Clearly something began happening in the late 1950s that significantly increased the advantage of incumbency.

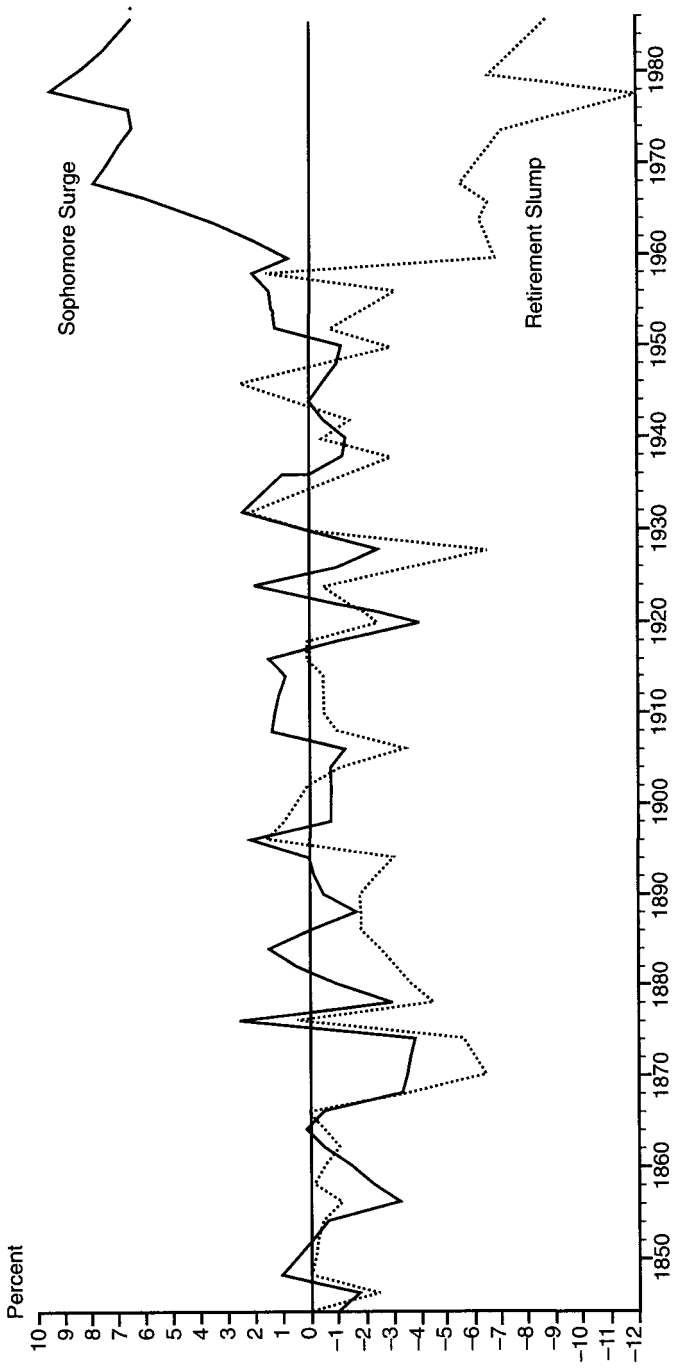
Alexander's early observation in 1915 on the potential role of congressional staff points to the essential element that has institutionalized the advantages of incumbency. The ability of professional staff to run what are essentially full-time reelection campaigns is the necessary element in the rise of all of the subsequent advantages of office holding. Year-round fundraising, "favors" for constituents, management of direct mail, and the promotion of constituent-oriented legislation are all made possible by staffing. They are also all unavailable to challengers.

The Legislative Reorganization Act of 1946 was the backhoe that began the entrenching of incumbency by initiating an unprecedented expansion of congressional staff. Congress justified the explosive expansion of its personnel as a counter to the growth of the Executive Branch, and to address an increasing congressional workload. Yet no matter how much bigger congressional staff became, it somehow

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

FIGURE 2

U.S. HOUSE ELECTION WITH MAJOR PARTY OPPOSITION, 1846–1986



NOTE: No data shown for election years ending in "2" and "4".
SOURCE: John R. Alford and David W. Brady (1991: 23).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

never caught up to its "workload," because new tasks that promoted reelection could always be found. House personal staff has grown enormously since the Legislative Reorganization Act. According to Norman Ornstein (1994: 126–27) the House employed 1,440 personal staff in 1947, 2,441 in 1957, 4,055 in 1967, 6,942 in 1977, and peaked at 7,584 in 1987. As a function of image making based on constituent services, staff has increasingly been moved closer to the voters. While in the late 1950s many congressmen had virtually no district staff, by 1992 almost one-third of all personal staff worked outside of Washington (Ornstein 1994: 131). And to confirm Alexander's warning, use of the franking privilege as a means of image making has increased from 45.6 million pieces of mail sent out in 1954, to regularly over 500 million pieces today (Ornstein 1994: 163).

Large staffs also run year-round campaign fundraising organizations, which they have done with increasing effectiveness. Although money has always been important in politics, it was never more so than in an age of modern media and incumbent image-building. Testifying before the Senate Subcommittee on Privileges and Elections in 1955, FCC Commissioner Freida Hennock noted,

> A major change has taken place in the process of elections since television has been the most effective means of campaigning. . . .The cost of reaching the public over television and radio at election time are so great that there is grave danger that the outcome of our elections will depend upon the war chests of the candidate and his party, and not upon the intrinsic worth of their qualifications.

Unfortunately, it is impossible to track campaign finances before 1974, when complete disclosure became mandatory. However the entrenching of incumbency seems to have coincided with the rise of television. Media scholar Robert Lichter (Lichter et al. 1986: 7) marks 1958 as the year "the age of television began," when the number of televisions approximately equalled the number of American homes. Perhaps not coincidentally, 1958 is also the precise year that the power of incumbency took off, according to Alvord and Brady's analysis. By 1960, the year of the first televised presidential debate, between John Kennedy and Richard Nixon, television was clearly having a powerful political effect. And if television is the engine to reelection, money is the fuel. With a full-time fundraising staff, incumbents have long had an advantage when it comes to building campaign war chests (Jacobson 1980, Malbin 1984). Television both decreased the unit cost of reaching voters, and provided the political process with a medium that was revolutionary in terms of its capacity to create public images. It maximized the impact of campaign funds by making possible, like never before, a personal appeal to voters.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Incumbents have dramatically expanded their capability to raise campaign funds, while challengers have clearly not been able to keep pace. Recently, PACs have provided decisive financial advantages almost exclusively to incumbents. In a circular arrangement, PACs today give to incumbents because incumbents win. Incumbents win largely because PACs give to incumbents. Party affiliation or ideology are practically irrelevant to the vast majority of PACs, who represent various economic interests. In 1976 House incumbents, on average, outspent challengers by $79,398 to $50,795. By 1992, the gap had widened to where incumbents spent $594,729 as compared with $167,891 for challengers (Ornstein 1994: 75). To illustrate incumbent advantage another way, the average amount of money spent on the congressional franking privilege during a single House term is about as much as the average challenger spends on an entire campaign.

Although the story of continued entrenched incumbency has been buried under news of the 1994 Republican "revolution," the recent election was, in fact, a conservative affair in terms of congressional turnover. The overall reelection rate in 1994 is consistent with four other elections in the era of entrenched incumbency. The U.S. House experienced a turnover of 20.8 percent in 1964, 21.1 percent in 1974, 18.6 percent in 1982, 25.3 percent in 1992, and now in 1994, 20 percent. Likewise, 91.3 percent of all incumbents who ran in 1994 defeated their challengers, a figure consistent with other elections in the era of entrenched incumbency.

If the advantage of incumbency is at least 7 percent, as Alvord and Brady suggest, then the power of incumbency disguised the true scale of popular sentiment in favor of the Republicans over the Democrats in 1994. If each challenger had gained 7 percent at the expense of the incumbent in the last election, then the Republicans would have won between 32 and 41 additional seats. Indeed, the "gridlock" experienced in recent times appears to be a function of Congress's lack of responsiveness to changing moods in the electorate, especially as compared with the relative responsiveness of the Executive Branch. Entrenched incumbency shortens presidential coattails, and retards the force of popular sentiment on Congress.[3]

The Reforming Impulse

Although the advantages of incumbency have not gone unnoticed, most reforms proposed fall far short of addressing the problem. Not

[3]If House elections do not accurately reflect the mood of the electorate, then members of Congress may not necessarily represent the beliefs of their constituents either. For more on this, see Robert A. Bernstein (1989).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

unexpectedly, Congress itself has not seriously considered any reform that would affect the system in a meaningful way. A host of proposals addressing the subject of campaign finance are regularly kicked around by public interest groups. They include various degrees of PAC reform, public campaign financing, and channeling campaign contributions through political parties. There have also been proposals to further restrict franked mail beyond the minimal reforms that Congress has already enacted. In addition, Freshman Republicans recently tried to implement modest cuts in House personal staff, but were thwarted by the leadership. So far, the 104th Congress has restricted cuts in personnel to the staff of abolished committees. Not only have most meaningful reforms failed to gather any significant degree of political momentum, but most only address a single aspect of incumbent power. Term limits is not only the sole reform with popular political clout behind it, but it also provides a comprehensive solution by guaranteeing turnover, ensuring periodic competitive open-seat races, and undermining cravings for perpetual reelections. By comparison, all other proposals, taken individually, only tinker with a system that requires structural reform.

Members of Congress have sensed the political power of the term-limit debate. Republicans have exploited the popularity of the issue by incorporating term limits in their "Contract with America." However, a majority of House Republicans who claim to support term limits have also tried to water down the reform by allowing six elections for members of the House. By contrast, only 3 of the 22 states that passed term-limit laws would have allowed such generous tenure for their federal representatives. Even the watered-down version failed to garner enough votes to send a constitutional amendment on to the states.

Permitting as many as five reelections for members of the House would both keep its members susceptible to the narrow and short-term interest pressures that fuel reelections, and also insulate that body from popular mandates. Only strict term limits will tend to make the U.S. House more immediately reflective of the American people and discourage faction, or special interests, as Madison intended.

Indeed, Madison had proposed a one-term limit for members of the U.S. House when he designed the Virginia Plan, as if he understood that open-seat elections most accurately reflect the immediate will of the people on their government (Erickson 1995). Madison and the other founders knew that special interests, especially economic interests, would always try to manipulate the political system, and that officeholders might use their positions to secure their own reelections by serving special interests. The challenge was to create a political structure that discouraged such manipulation. For all of their vision,

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the founders could not have foreseen the rise of an industrial democracy, or the regulatory state and technological innovations that have accompanied it.

In fact, rotation of offices operated naturally during the first 100 years of the republic's history. During the progressive era, however, reformers lamented the growth of corrupt urban centers, and were stunned to discover the extent to which big business influenced government (McCormick 1981). They responded by trying to protect the old rural America they had known and idealized. In so doing they gave government new authority to manage the frightening new world in which they found themselves, but the power awarded to government was also necessarily placed in the hands of legislators. No new limits were placed on the people's representatives to correspond with the radically expanded role of government that arose out of the progressive era, the New Deal, and the Great Society.

Since about 1958, professional politicians have learned how to use the power of incumbency to perpetuate their tenure in office. In response, millions of American citizens, motivated by the "intuitive impulse" that assumes political careerism is inheritently dangerous, passed term-limit laws in every state where the issue came up for a vote. This historic popular reform effort was, however, brought to a screeching halt in the spring of 1995. Ruling against the constitutionality of state-mandated term limits in *U.S. Term Limits v. Thornton*, the Supreme Court invalidated the state laws in a decision with possibly profound consequences. In its wisdom, the High Court has left the fate of term limits legislation largely in the hands of those who benefit most from the status quo, professional politicians in the U.S. Congress. Given the unlikelihood of a Constitutional Convention, the vast majority of Americans who perceive the need for term limits must now rely on members of Congress to legislate against their own immediate self-interests, a very rare thing in politics. Whether or not Congress finally passes on to the states a constitutional amendment for meaningful term limits will prove to be a crucial test of the ultimate health of American democracy.

References

Adams, W.P. (1980) *The First American Constitutions*. Chapel Hill: University of North Carolina Press.

Alexander, de A.S. (1970) *History and Procedure of the House of Representatives*. New York: Ben Franklin.

Alford, J.R., and Brady, D.W. (1991) "Personal and Partisan Advantage in U.S. Congressional History, 1846–1986." In Silbey, J.H. (ed.) *The United States Congress: The Electoral Connection, 1789–1989*, vol. 1, 23–40. Brooklyn, N.Y.: Carlson Publishing.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Aristotle (1984) *Politics*. In *The Complete Works of Aristotle,* Book IV, Sec. 2. Edited by Jonathan Banes. Princeton: Princeton University Press.
Bailyn, B. (1967) *Ideological Origins of the American Revolution.* Cambridge, Mass.: Belknap Press.
Bensal, R.F. (1984) *Sectionalism in American Political Development, 1880–1980.* Madison: University of Wisconsin Press.
Bernstein, R.A. (1989) *Elections, Representation, and Congressional Voting Behavior: The Myth of Constituency Control.*, N.J.: Prentice Hall.
Brady, D.W., and Grofman, B. (1991) "Sectional Differences in Partisan Bias and Electoral Responsiveness in U.S. House Elections, 1850–1980." *British Journal of Political Science* 21(2): 247–56.
Cain, B.E.; Ferejohn, J.A.; and Fiorina, M.P. (1987) *The Personal Vote: Constituency Service and Electoral Independence*. Cambridge, Mass.: Harvard University Press.
Congressional Quarterly Guide to U.S. Elections. (1985) 2nd ed. Washington, D.C.: Congressional Quarterly.
De Tocqueville, A. ([1835] 1965) *Democracy in America* (2 vols.) In *Great Books of the Western World*. Edited by M.J. Adler. Chicago: Encyclopaedia Britannica.
Dixon, R.G. (1968) *Democratic Representation: Reapportionment in Law and Politics*. New York: Oxford University Press.
Erickson, S.C. (1993) "A Bulwark Against Faction: James Madison's Case for Term Limits." *Policy Review* (Winter): 76–78.
Erickson, S.C. (1995) "James Madison, a Term Limits Radical: The Father of the Constitution's Solution to Congressional Gridlock." U.S Term Limits *Outlook Series*, Washington D.C.
Erikson, R.S. (1971) "The Advantages of Incumbency in Congressional Elections." *Polity* 3: 395–405.
The Federalist Papers ([1787–88] 1976) Edited by L. DeKoster. Grand Rapids, Mich.: William B. Eerdmans.
Fiorina, M.P. (1977) *Congress, Keystone of the Washington Establishment*. New Haven, Conn.: Yale University Press.
Grantham, D.W. (1988) *The Life and Death of the Solid South*. Lexington: University of Kentucky Press.
Harrington, J. ([1656] 1977) *The Commonwealth of Oceana* and *The Prerogative of Popular Government*. In *The Political Works of James Harrington*. Edited by J.G.A. Pocock. Cambridge: Cambridge University Press.
Hennock, F.B. (1955) Testimony before the Subcommittee on Privileges and Elections Regarding the Federal Elections Act of 1955. 84th Cong., 1st sess.
Jacobson, G. (1980) *Money in Congressional Elections.* New Haven, Conn.: Yale University Press.
Kernell, S. "Toward Understanding Nineteenth Century Congressional Careers." *American Journal of Political Science* 21: 669–93.
Kleppner, P. (1987) *Continuity and Change in Electoral Politics, 1893–1928.* New York: Greenwood Press.
Kraditor, A. (1965) *The Ideas of the Women's Suffrage Movement, 1890–1920.* New York: Columbia University Press.
Lichter, S.R.; Rothman, S.; Lichter, L. (1986) *The Media Elite: America's New Power Brokers.* Bethesda, Md.: Abler and Abler.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Madison, J. (1962) *The Papers of James Madison*. Edited by W.T. Hutchinson, W.M.E. Rachal, and R.A. Rutland. Chicago: University of Chicago Press.
Malbin, M. (1980) *Unelected Representatives*. New York: Basic Books.
Mayhew, D.R. (1974a) "Congressional Elections: The Case of the Vanishing Marginals." *Polity* 6: 295–317.
Mayhew, D.R. (1974b) *Congress: the Electoral Connection*. New Haven, Conn.: Yale University Press.
McCormick, R.L. (1981) "The Discovery that Business Corrupts Politics: A Reappraisal of the Origins of Progressivism." *American Historical Review* 86: 247–74.
McCormick, R.L. (1986) *The Party Period and Public Policy: American Politics from the Age of Jackson to the Progressive Era*. New York: Oxford University Press.
MacNeil, N. (1963) *Forge of Democracy: The House of Representatives*. New York: David McKay.
Ornstein, N.J.; Mann, T.E.; and Malbinn, M. (eds.) (1994) *Vital Statistics on Congress, 1993–1994*. Washington, D.C.: Congressional Quarterly.
Pocock, J.G.A. (1975) *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton, N.J.: Princeton University Press.
Richardson, J.D. (ed.) (1896) *A Compilation of the Messages of the Presidents*, 10 vols. Washington D.C.: U.S. Government Printing Office.
Richmond Enquirer, 8 November 1822.
Robbins, C. (1968) *The Eighteenth Century Commonwealthman*. New York: Athenaeum.
Struble, R., Jr. (1980) "House Turnover and the Principle of Rotation." *Political Science Quarterly* 94 (Winter): 649–67.
Trenchard, J., and Gordon, T. ([1755] 1995) *Cato's Letters*. Edited by R. Hamowy. Indianapolis, Ind.: Liberty Classics. (The essays were first published in London from 1720 to 1723.)
Wiebe, R. (1967) *The Search for Order, 1877–1920*. New York: Hill and Wang.
Wiebe, R. (1984) *The Opening of American Society: From the Adoption of the Constitution to the Eve of Disunion*. New York: Knopf.
Wood, G. (1969) *The Creation of the American Republic, 1776–1787*. Chapel Hill: University of North Carolina Press.
Wood, G. (1993) *The Radicalism of the American Revolution*. New York: Vintage Books.
Woodward, V. (1955) *The Strange Career of Jim Crow*. New York: Oxford University Press.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Competitive Markets for Electricity Generation

John C. Moorhouse

The Trend toward Competition

This paper discusses the feasibility of replacing regulation or state ownership with market competition in electric power generation. Interest in competitive electricity markets has been stimulated by recent experiences in the United Kingdom and Brazil with privatization, with deregulation in the United States, and with partial reform in Norway. While the Thatcher government policy offers lessons about the process of privatizing a state monopoly enterprise, the American experience becomes relevant to our understanding of how increasingly competitive markets for electricity actually work. The continuing growth of competition in American electricity markets is an unanticipated consequence of the 1978 passage of the Public Utility Regulatory Polices Act. Designed as a conservation measure, PURPA established the right of cogenerators and Independent Power Producers (IPPs) to sell electricity to local regulated Investor-Owned Utilities (IOUs).

Such rights were broadened substantially by the passage of the Energy Policy Act of 1992 which requires transmission line owners to wheel bulk power (Walters and Smith 1993). Thus, under current federal regulations nonutility power producers can sell electricity to any utility on the grid. Furthermore, in April 1994, the California Public Utility Commission adopted a policy establishing complete open access to all power producers. By 1996 independent generators can compete to sell electricity directly to large industrial customers, effectively bypassing traditional utilities. By 2002, the policy permits all electricity consumers, regardless of size, to purchase electricity

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

The author is Professor of Economics at Wake Forest University. A shorter version of this paper was published in 1995 in French as "Risque et Concurrence," *Le Communicateur* 28: 199–223. The author would like to thank Henri Lepage for his helpful comments on an earlier draft.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

from *any* utility or independent generator on the grid. No longer will the consumer be restricted to buying electricity from the local utility. A competitive market for generation will have been established (Hoffman 1994).

There are further lessons to be learned from the way the natural gas market has developed in the United States following deregulation. Particularly relevant have been the separation of production, pipeline transportation, and distribution into independent operations; the growth of the spot and futures market for gas, and the latter's replacement of the long-term contract in the buying and selling of natural gas (Doane and Spulber 1994). In the United States, the majority of natural gas sales to local distributors are on the spot market (Lyons 1990).

The system evolving in the United States provides increasing competition and diversity among generators. They vary from established utilities, IPPs, and cogenerators to small producers that use renewable fuels and other nonutility generators. In 1983, the NUGs provided 2.5 percent of total United States generating capacity, by 1991 they provided 9 percent of the total and were building over half the new capacity installed in the nation (Michaels 1993). Once the genie was out of the lamp, the benefits of competition insured that more open markets for generation would spread.

The European Community is addressing these same issues and has agreed to draft directives calling for open access in energy markets. As of January 1993, the European Commission seeks to let large users of electricity—those using 100 gigawatts or more of power per annum (aluminum, steel, chemicals, glass, and fertilizer)—to purchase electricity from any supplier in the Community. Energy Commissioner, Antonio Cardos e Cunha (in the *Financial Times* 1992), summarizes the goal,

> Our aim is to transform the energy market in Europe—which is fundamentally national and based on administrative focusing on price—and replace it with a European unit in which cross-border trades could be significant and prices would react according to negotiations between buyer and seller.

The paper is organized as follows. The next section deals with the twin ideas that the generation, transmission, and distribution of electricity is a natural monopoly and involves only technical issues. The following three sections present a model of competitive generation, outline the major advantages of competitive electricity generation, and discuss how electricity would be priced in a competitive market. The next section compares and contrasts how the market addresses

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the problem of risk with the way it is dealt with under regulation or state ownership. The final section presents a summary and conclusion.

A Change in the Climate of Ideas

The British and American reforms arose in part because two fundamental ideas have been successfully challenged during the last two decades. The first idea is that the electricity industry is a natural monopoly. The second, independent but complementary, idea, particularly influential in Europe, is that the generation, transmission, and distribution of electricity raise purely technical problems to be solved by engineers.

Ideas have consequences. Any discussion of the electricity industry always starts with the observation that the industry is capital intensive and, because power cannot be stored, it must be generated so as to meet instantaneous demand. These characteristics coupled with the natural monopoly argument are said to justify government franchising and regulation or, alternatively, state ownership.[1]

The notion that the provision of electricity presents only technical issues naturally justifies the vertical integration of generation, transmission, and distribution into a state-owned monopoly. It is worthwhile exploring the challenge to these two ideas in greater detail because opening up the debate has created an intellectual climate more conducive to reform.

Natural Monopoly

The traditional economic justification for the regulation or state ownership of electric utilities is that utilities are natural monopolies. Economies of scale and scope, and the economies associated with vertical integration mean that unit costs decline throughout the relevant range of production as output increases. Such economies preclude competition, according to the conventional view, because a single firm could supply the entire service area at lower cost than could two or more firms. Given its cost structure, an established utility could undercut its rivals and drive them from the market. Moreover, attempted entry represents a waste of resources either because of an unnecessary duplication of facilities or because such investment would not be viable in the face of undercutting. Secure from competition, the monopolist would exploit the consumer if not for regulation or state ownership.

[1]The latter never proved popular in the United States except for small municipal distribution companies and two huge Federal power facilities, TVA and Bonneville.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

During the last 20 years, a growing body of evidence, primarily from the United States, has undermined the view that electric utilities are natural monopolies (Hammond 1986). Leonard W. Weiss (1975: 136) argues that the postwar technological trend toward building larger generating plants, in the United States, ended in the mid-1960s, "Most important regions could support enough generating plants to permit extensive competition if the plants were under separate ownership and had equal access to transmission and distribution."

In the late 1960s the conventional wisdom in the industry was that new plants burning fossil fuels had to have a generating capacity of from 1,000 to 1,500 megawatts and that the optimal capacity of a nuclear plant might be even larger. As Edward Berlin, Charles J. Cicchetti, and William J. Gillen (1974: 9) observe, thermal efficiency is not the whole story; reliability plays a role:

> The effect of unit size on reserve requirements is a straightforward problem. The contingency to be guarded against is that a particular unit will not be available when needed. If generating capacity consists of a large number of small units, risk is spread over each of the units. . . . The forced outage rate for fossil-fueled plants over 600 mw is more than twice that of plants below 600 mw.

Within bounds, the implication is clear: smaller plants are more reliable than large-scale plants. When dynamic reliability is played off against static design efficiency (economies of scale and scope), the optimal scale for plants is smaller than originally thought. Finally, a number of studies suggest that economies of scale and scope in the generation of electricity may not be significant enough to undergird natural monopoly.[2]

William J. Baumol, John C. Panzar, and Robert D. Willig (1982) argue, in their seminal work *Contestable Markets and the Theory of Market Structure*, that even if economies of scale and scope are present, they are neither necessary nor sufficient for natural monopoly to exist. Making the critical dynamic distinction between fixed costs (costs that are not zero when output is zero) and sunk costs (capital invested in plant and equipment that has no alternative use or opportunity cost), they observe that only sunk costs give an existing firm the cost advantage necessary to insulate it from competition. Absent significant sunk costs, potential entry undermines monopoly pricing; natural monopoly is not sustainable. Such markets are said to be contestable (Panzar and Willig 1977).

[2]See Christensen and Green 1976, Huettner and Landon 1978, Cowing and Smith 1978, Stewart 1979, and Jaskow and Rose 1985.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

While it is not obvious that electric markets can legitimately be classified as contestable, multiproduct production, uncertain demand, risk aversion, and interproduct (interfuel) substitution are all conducive to potential entry (Sharkey 1982). Don Coursey, R. Mark Isaac, and Vernon L. Smith (1984: 111) conclude from their research that

> the most significant result . . . is that the behavioral predictions of the contestable market hypothesis are fundamentally correct. It is simply not true that monopoly pricing is a 'natural' result of a market merely because firms in the market exhibit decreasing costs and demand is sufficient to support no more than a single firm.

Thus this research suggests that competition can be substituted for government regulation and state ownership to assure good performance in, at least, the generation of electricity.

Technological Efficiency

The other intellectual tradition that has been influential in the governance of electric utilities argues that the generation, transmission, and distribution of electricity can be reduced to a set of purely technological problems. Because engineering efficiency should guide the design and development of electric power facilities, this view concludes that engineers should operate such facilities as state enterprises. The reliable provision of electricity is vital to public welfare. The public is best served when a staff of civil servants, trained in engineering, attend to these technical matters.[3]

Yet the generation and distribution of electricity is not primarily, let alone exclusively, a matter of technology. Consider the concept of efficiency. Technological efficiency is always defined as the rate at which an input is converted into an output; a ratio of two quantities. For example, one definition of efficiency is the ratio of work done to energy input. The ratio is less than one, because some energy is dissipated in the form of waste heat during the process that converts energy into work. A process is said to be more efficient than another (see Heyne 1987: chap. 6) if, in the first process, a higher proportion of potential energy is converted into kinetic energy than in the second or, equivalently, if the first process generates less heat loss than the second process.

On that definition, the most technologically efficient means of producing electricity would be the method that converted the highest proportion of the potential energy of its fuel into electrical energy.

[3]In a fascinating discussion, Friedrich A. Hayek (1964) traces the intellectual roots of this approach to policy engineering to the *Ecole Polytechnique* (see his *The Counter-Revolution of Science,* especially Part 1, chap. X, and Part 2, chap. IV.).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

By that criterion, nuclear plants are more efficient than coal-fired plants, and coal-fired plants are more efficient than are hydrogenerating stations.

If electricity were the only commodity of economic value, the technological and economic concept of efficiency would come down to the same thing. But of course, literally tens of thousands of goods and services are valuable. The inputs, including fuel, that are necessary to generate electricity are scarce resources in the sense that they have many alternative uses—alternative uses that also satisfy human wants. Thus, the most efficient way to produce electricity, from society's point of view, is the method that maximizes the difference between the value of output and the value of inputs.

The objective data of the engineer (conversion rates) is only one factor in the determination of the subjective values that measure economic efficiency. Hydroelectric power plants, the least technologically efficient plants, may be the most economically efficient means of producing electricity when the value of the resources consumed in production and the environmental consequences of hydropower are taken into account.

Writing 70 years ago, the engineer Otto Goldman (1923: 84), using different terminology, makes the same argument,

> It seems peculiar and indeed is unfortunate that so many authors in their engineering books give little, or very little, consideration to costs in spite of the fact that the primary duty of the engineer is to consider costs in order to obtain real economy—to get the most power, for example, not from the least pounds of steam, but from the least number of dollars and cents: to get the best financial efficiency.

Absent technological change, objective engineering data on energy conversion rates remain constant and provide an unambiguous ranking of the different methods for producing electricity according to technological efficiency. By contrast, the ranking of production methods according to economic efficiency can change as the result of either a technological innovation or a change in the relative value of inputs. An example of the latter would arise if the market valuation of nuclear fuel rose and the price of coal dropped. Under such circumstances, the value of the resources consumed by coal-fired plants might drop below that of nuclear plants, per unit of electricity produced, and coal-fired plants would become more economically efficient than nuclear plants.

A shift in subjective valuation changes the measure of economic efficiency while having no effect on the measure of technological efficiency. The twin measures of efficiency are fundamentally different. In terms of human welfare, the most efficient means of production

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

is always gauged in terms of economic efficiency, the ratio of the subjective value of output to the subjective value of inputs.

The concepts of engineering and economic efficiency diverge for two reasons. First, there are different methods of producing electricity, each of which uses a different combination of inputs. These inputs are scarce; they have alternative uses. Which use is the more valuable depends on subjective assessments revealed only by market exchange. Second, electricity is not a single good but a multiproduct having many characteristics including peak and off-peak timing of production and consumption, and reliability of service. A method of production that is efficient for producing a base load may be inefficient for meeting peak-load demands. Narrow technical efficiency provides no meaningful criterion for ranking production methods when inputs have alternative uses or when production involves complex goods with multiple characteristics. Electricity is just such a good. The higher dimensionality of the concept of economic efficiency and the greater information costs of measuring economic efficiency present a problem that is best solved by voluntary market exchange.

What makes the notion of objective technological efficiency so congenial to the idea of state ownership and operation of this vital industry is that it avoids the necessity of obtaining and acting on information about market valuation. Technological efficiency seems to point to *the way* of generating and distributing electricity. All the information necessary to do so is known by the civil servants acting on behalf of the public interest.

The natural monopoly thesis and the notion that producing electricity is a purely technical matter are powerful ideas that die hard; they have shaped the structure and governance of the electricity industry throughout this century. This paper argues that, to the contrary, the efficient use of resources to produce and distribute electricity requires information about relative valuations that can be generated only by market exchange.

Competitive Generation

Theoretical arguments, statistical evidence, and—of increasing importance—experience favor the viability of competitive markets for the generation of electricity. In general, reform means the substitution of market competition for state ownership or government regulation to ensure good performance. The nature of the particular reform necessary to introduce competition into the electricity industry depends, of course, on the current structure of the industry. If the industry is a state-owned monopoly, as, for example, in the case of

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Electricité de France (EDF), then reform means full privatization and the separation of generation from transmission and distribution functions. If the industry consists of investor-owned utilities, subject to government regulation, as in the United States, then deregulation of generation capacity coupled with the separation of generation from transmission and distribution is required.

Elements of Reform

Thoroughgoing reform has four major elements: (1) private ownership of electricity industry facilities; (2) open access for generators to transmission facilities; (3) a minimum of three independently owned generating stations that could potentially compete for consumers within each regional electricity market or service area;[4] and (4) separation of generation from transmission and distribution.[5] More radical reform would be difficult because privatization, restructuring, and complete deregulation of transmission and distribution facilities remain controversial.

By contrast, decentralized, competitive markets at the generation stage are not only conceptually desirable, but are becoming increasingly important in the United States. John Tschirhart (1991: 27) observes, "Entry of new firms into the electric power industry is becoming commonplace. The entrants typically are unregulated firms that compete with regulated electric utilities only in the generation stage of the latter's integrated structure." By 1990, a decade after the reform movement got under way in the United States, cogenerators and unregulated IPPs were building more generating capacity than were traditional utilities. For example, Southern California Edison buys 30 percent of its power from NUGs. The Midland Co-Generation Project in Michigan consists of 12 gas turbines with a generating capacity of 1,343 megawatts. In New York State, construction of new generating capacity is determined by competitive bid with many contracts being awarded to NUGs (Tschirhart 1991). The Alamita Company, in Arizona, is an independent power company that sells electricity to bulk customers, Tucson Electric Power Company, and Southern California Edison (Herriott 1989a).

[4]Producers need not be located within the region, but only interconnected to the market via transmission lines. Independent producers, in some instances, could be created by breaking up existing generating capacity into a (relatively small) number of companies that would compete to sell electricity. Other competitors in a regional market could include IPPs and cogenerators.

[5]One approach is to treat the transmission grid as a regulated common carrier providing access and central dispatch to all power producers. Distribution companies could be regulated as regional (private) monopolies.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Compared with the deregulation of IOUs, privatization of a state-owned monopoly requires a complete and fundamental change in the structure of ownership and property rights in the electricity supply industry in order to obtain the benefits of increased efficiency and innovation. A shift from public to private ownership refocuses the goal of the producer toward profits. Pursuit of the latter provides a strong economic incentive, in a competitive environment, to improve and maintain the quality of customers services, monitor costs more closely, and invest in productivity-enhancing technologies. These incentives are blunted by state ownership. With respect to privatization, the British experience since 1989 seems more germane than does the regulatory reform the United States has been undergoing since 1978.

The Advantages of Competitive Generation

Competitive generation envisions a market within which independent firms compete on the basis of price to sell electricity directly to large industrial customers (bulk wheeling), and to supply electricity, via common carrier transmission, to distributors who in turn sell power to final users (Rozels 1989, Bushnell and Oren 1994). Producers may specialize or diversify by load characteristic. For example, some may prefer to compete for long-term base-load contracts. These firms are likely to own hydro and nuclear power plants. On the other hand, firms with fossil fuel plants might seek to supply base and cycling loads. Finally, producers with gas combustion turbines and cogenerators could compete to meet peak loads. Other firms may diversify and be ready to compete for base, cycling, and peak loads. Peak and off-peak loads are defined by day, week, and season.

Prices charged for each type of service (peak and off-peak load, daily to seasonal) could be established by contract, 24-hour advance notice, and in spot markets. Unit prices could vary by the amount of electricity purchased per period (Wilson 1993). As a result, customers would face more service options and a more complex pricing scheme.

There are a number of advantages to having a variety of types of generators linked to the transmission grid. The first major advantage involves cost savings. At any given moment, power is supplied to the transmission grid by the firm with the lowest marginal costs. Dispatch according to merit (from lowest to highest marginal cost including line loss[6]) saves resources and reduces the cost of generating electricity.

[6]Typical line loss approximates 1 percent per 100 kilometers. Line loss limits the extent of the market, because the loss rate is a nonlinear increasing function of distance.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Because the different plants may have different load characteristics, peak and load duration curves, generating capacity can be more fully utilized and additional capital resources saved.

Moreover, the variety of generating equipment and the larger number of independent producers adds diversity to the system, lowering the probability of widespread equipment failure, and, thereby, reducing the amount of excess capacity required to provide a given level of service reliability. The availability of electricity from alternative suppliers means that generators can coordinate routine plant maintenance. The added flexibility in scheduling repairs lowers plant maintenance costs. These four sources of cost savings mean lower prices to the consumers of electricity (Herriott 1989b).

The second advantage of competitive generation is that a spot market for electricity will develop. The ability to sell electricity on the spot market increases the generator's flexibility in scheduling production. Consider the position of the monopoly producer. The firm must wait for customers to "flip the switch" and demand electricity. In order to meet fluctuating demand, the monopolist must maintain costly "spinning" capacity as a reserve. By contrast, competitive generators can sharply reduce their idle spinning capacity and simply schedule efficient production. They need not wait for customers because they can sell power on the spot market at any time. Efficient scheduling reduces production costs. The dynamism and flexibility of the competitive market is the antithesis of the passive waiting game that the monopolist must play.

Moreover, the presence of a spot market means that less idle capacity must be maintained in order to provide a given level of service reliability. Shortfalls and emergencies can be met by purchasing power on the spot market. Demand and supply are equilibrated by flexible spot prices.

The third advantage of competitive generation is that the market will provide an array of service standards that more closely match the mosaic of consumer preferences (Caves, Herriger, and Kuester 1989). It is a shibboleth of the industry that a utility must stand ready to supply electricity to all demanders at all times. But of course such a standard of reliability is met and maintained only at high cost. Vertically integrated utilities invest in expensive excess (off-line) capacity to ensure a high degree of reliability. For example, gas combustion turbine plants make up much of the spinning standby capacity in the United States.

Not all consumers seek such a high quality of service all of the time. With the option of choosing among competitive generators, distributors have an incentive to discover the level of reliability

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

demanded by different consumers. Consumers could be offered priority service with a schedule of electricity rates increasing with the level of reliability. Customers upon whom infrequent but uncertain service interruptions impose relatively low costs could save (with certainty) by subscribing to lower priced, lower priority service. Conversely, customers for whom interruptions impose high costs would subscribe to higher priced, more reliable service (Smith 1989).

Interruptible service does not mean customers are without service altogether. Rather interruptions involve a reduction in power. Typically in the United States, for example, interruptible residential service means that the consumer's heavy appliances (air conditioner, hot water heater, or space heater) are turned off when the utility sends an electrical impulse that trips a circuit breaker. Full power is restored after the emergency or peak load has passed. In other systems, customers purchase a fuse that guarantees a minimum level of power not subject to curtailment; power use above that level can be interrupted. The higher the minimum level of service chosen by the consumer the higher the price of electricity (Wilson 1989, Woo 1990).

By offering such choices, distributors obtain information about consumer preferences for reliability. With this information distributors know, for instance, how much and at what prices electricity should be purchased on spot markets during peak periods or emergency outages.

In spite of costly excess capacity, outages occur. Unexpected equipment failure or peak demands temporarily greater than generating capacity cause blackouts or rolling brownouts. The resulting random rationing is inefficient because it represents an arbitrary distribution of an unpriced resource—reliable service. By contrast, during an emergency shortage with priority service, consumer service is interrupted in order of ascending priority, from lowest to highest reliability. Because consumers have revealed their preference for service reliability, a priority system reduces the aggregate cost to consumers of interruptions, compared with random rationing.

According to Hung-Po Chao and Robert Wilson (1987), priority service offers significant efficiency gains over random rationing with fixed electricity rates. The authors argue that advances in the microelectronic technologies of metering and control make selective, ordered rationing economically feasible. Similarly, Abraham Grosfeld-Nir and Asher Tishler (1993) found variations in outage costs for a sample of Israeli industries, which suggested to them that there is a potential for large savings if priority service is offered.

Priority service reduces costs in three ways. First, it reduces the excess capacity carried in order to provide the reliability demanded by consumers. Second, when outages occur and service is rationed,

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

priority service minimizes the costs of interruptions borne by consumers. Third, because consumers have revealed their preferences for both the quantity and reliability of electricity service, long-run capital planning is more efficient, hence more cost effective.

Priority service can be fine-tuned by offering rate schedules based on both the frequency and average duration of service interruption, and on whether service would be interrupted only upon prior notification. Those willing to bear the risk of uncertain interruption would pay lower electricity rates. Finally, as Shmuel S. Oren and Joseph A. Doucet (1990) suggest, suppliers can provide interruption insurance with premiums based on the probability of rationing and compensation specified.

A competitive market in electricity generation would offer a much broader array of services than do state monopolies or regulated generators. Perhaps it is not surprising that 70 percent of United States private utilities, facing new competitive pressure at the generation stage, now offer some form of voluntary interruptible service (Strauss and Oren 1993).

Such services are less likely to be offered by either state monopolies or heavily regulated private utilities than by competitive generators precisely because priority service economizes on capital capacity. For a state monopoly, offering priority service would reduce the size of the enterprise and budget, the level of employment, and ultimately power within the bureaucracy. For private utilities operating under cost plus pricing regulation, more capital means higher accounting profits. Moreover, for both state and private monopolists system reliability is a powerful hook on which to hang a request for a larger budget or rate increases. Tailoring service standards to customer demands is the last thing to expect from a state or private monopoly.

The fourth advantage of competitive generation is innovation. Experience in the United States with deregulation of transportation and telecommunication and, on a more limited basis, the electric power industry suggests that competitive pressure tends to make industries more innovative. Competition not only leads firms to be more responsive to consumer demands, monitor costs more closely, and compete on the basis of price, it provides an incentive to be innovative because that may be the only way to get a temporary jump on rivals. Developing a new consumer service, a better method of reducing costs, or a faster way of dealing with problems promises the innovator a competitive edge. Because research and development is risky and offers little reward to state enterprises or utilities with monopoly franchises, evidence suggests that the latter organizations are less innovative.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Competitive Pricing of Electricity

Private producers of electricity seek to earn profits. If their operations are regulated by market competition rather than state regulations governing prices and rates of return, producers will monitor costs more closely at *all margins* and, further, offer electricity at prices determined by marginal costs. Because the cost of generating electricity varies by reliability of service and whether demand falls on a peak or off-peak period, we expect a more elaborate pricing system to emerge under competitive conditions.

Industrial Customers

With competitive generation, industrial customers can purchase electricity from the regional distributor or directly from a generator via bulk wheeling. Industrial users can shop for electricity. Service standards and prices would be negotiable and subject to long-term contract. Unanticipated requirements can be met on spot markets (Schweppe et al. 1988). Retail spot market prices for industrial users would equal wholesale spot prices (determined by competition among generators) plus transmission and distribution charges. The regulatory authority would guarantee open access for generators to the grid, approve transmission and distribution charges, and require that distributors base retail spot prices on prevailing wholesale spot prices.

It may be the case that few, if any, industrial or commercial customers, would want to purchase power on the spot market. Buying and selling electricity at volatile spot prices involves relatively high transaction costs and risk. The point remains that competitive generation makes retail spot markets viable. Also note that with competitive generation, an industrial customer who is buying electricity from a supplier at 11:00 a.m. may be selling electricity to that same supplier at 7:00 p.m. (Rose and McDonald 1991). In competitive markets, the roles of buyer and seller can change periodically.

Residential Consumers

Residential consumers too are in a position to contract with any producer on the network and thus can shop for the most desired type of service and prices. Residential consumers might pay set prices from a menu establishing time-of-day and time-of-year rates, and premiums and discounts based on reliability of service. Prices for interruptible service would vary depending on probability and expected duration of outages and whether or not prior notification is specified. Residential consumers are unlikely to want to purchase power on the spot market. Fluctuating spot prices would generate too much uncertainty for the

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

typical residential user. The vast majority of households are likely to opt for menu prices, based on service characteristics, in part, because such a pricing scheme makes estimating expenditures on electricity easier.

Distribution Companies

With competitive generation, distributors have many options. Indeed, distributors become the most important shoppers for electricity in the system. Depending on the quality of service demanded by their customers, distributors' degree of risk aversion, and their ability to diversify, distribution companies seem likely to purchase their base and expected peak loads by long-term contract with shortfalls and emergency requirements met in the spot market (Doucet 1994). For example, Florida Energy Brokers operate a wholesale spot market for electricity which posts prices hourly.

Transmission Services

An analysis of the pricing of transmission service is beyond the scope of this paper, but a few comments are in order. Private transmission companies would be regulated as common carriers that guarantee open access to all potential suppliers of electricity. The transmission company's role is pivotal as it performs central dispatch. Operating the grid involves establishing network policies and procedures, assigning load to each generating station, maintaining network lines, scheduling energy transfers, and accommodating bulk wheeling.

Transmission rates should be set so that (1) all transmission costs, discussed below, are covered; (2) short-run demand equals the short-run supply of electricity, and (3) price provides guidance for optimal investment in future transmission capacity.

The transmission of electricity is governed by certain physical laws. These should be interpreted as *technical constraints* within which the market determines the quantity and price of electricity supplied at any given time by any given generator. Thermal and voltage limits define transmission capacity (Hogan 1993). Moreover, electricity placed on the grid flows according to Kirchoff's laws—for the purposes of this paper, electricity flows along the path of least resistance. Thus, for example, electricity sold by Generator A to Industrial Customer B may not travel along the "contract path," that is, the shortest line within the network that directly links the buyer and seller. Depending on circumstances, electricity introduced into the network at any point may give rise to "loop flow" affecting all suppliers to the grid. Loop flow can disrupt the quality and reliability of service to everybody

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

taking electricity from the grid at the moment additional power is introduced.

While conceptually loop flow can be a problem, central dispatch is designed to direct traffic so as to avoid problems associated with these technical constraints. Linking a number of independent power producers increases the load diversity of the system and the number of energy transfers per period thus reducing the significance of loop flow (Walters and Smith 1993). As William W. Hogan (1992: 216) argues,

> When loop flow is a small part of power economics then informal swaps can balance out the effects over time, and when all parties are members of the same transmission club, it is reasonable to employ the contract path fiction as a practical accommodation in crafting power contracts.

Network rules would, of course, establish minimum technical standards for putting electricity onto the grid: long-term contracts setting forth transmission rights for each generator, and fees for transmission services. Transmission fees might consist of a two-part tariff: an access charge covering capital costs and a fee set equal to the marginal cost of providing transmission service. Because the latter includes maintenance and congestion costs, fees would vary with the volume and distance power is transferred. When appropriate, a higher congestion fee would be charged short-term users because of the higher transaction costs of arranging dispatch on short notice (Doyle and Maher 1992).

The Market Response to Risk

It is important to understand the market response to the increased risk associated with the introduction of competition into the market for generating electricity. Before discussing this issue, however, it is desirable to outline the sources of risk that are independent of market structure. The production and distribution of electricity is subject to predictable time-of-day and seasonal variations in demand and input prices. In addition, it is subject to stochastic equipment breakdown, unexpected fluctuations in demand that must be accommodated instantaneously, and unexpected changes in input prices, particularly those for fuel.

Typically a vertically integrated state monopoly deals with fluctuations in demand and random equipment failure by carrying excess capacity, including redundant backup capacity. It may also address predictable fluctuations in demand by offering peak-load pricing schemes, although the incentive to do so is weakened by state owner-

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ship or regulation. Installing excess capacity is more attractive to the bureaucracy. Peak-load pricing was well understood long before state or franchise monopolies implemented time-of-day service.

Competitive generation produces at least two additional sources of risk: a more complex pricing structure (discussed above), and loop flow problems when independent producers put electricity into the transmission network. But if decentralized markets introduce additional risk, they also provide a broad array of ways of dealing with it. All of these sources of risk potentially influence the quality of service to the final consumer of electricity. In general, the market offers methods to reduce risk and to price risk so that it can be spread or shared optimally.

Fluctuations in electricity prices and interruptions in service generate risk for the buyer and seller. In the case of price variations, the payer finds it more difficult to forecast expenditures and plan accordingly. The seller finds it more difficult to forecast gross receipts. Uncertain interruptions of service complicate consumer and producer planning. The risk borne by a given party depends on the probability of an untoward event (an outage), the magnitude of the event (extent and duration of an outage or the magnitude of an unexpected price change), the costs imposed by the event, and the degree of risk aversion of the affected party. Because reducing risk, transferring risk to parties willing to bear it, and spreading risk among all parties are valuable activities, it is not surprising that the market provides for all three.

Consider how a generator faces the risk of uncertain prices for electricity. First, the producer can sell power by long-term contract to large industrial customers and regional distributors. Contracts specify prices and adjustment clauses. Thus, only a small proportion of its output may even be exposed to unknown price fluctuations (Jaskow 1988). Second, selling on the spot market on a regular basis offers normal returns because prices regress toward the mean over a large number of sales. By selling regularly in the spot market, the producer is reducing risk through diversification. Third, the producer can hedge spot market sales in the futures market.

The futures market for electricity is really a market for future generating capacity, not future electricity. The generator is said to be "long on capacity"; therefore, the firm hedges by selling futures contracts for the amounts and periods corresponding to its desired sales profile. On the other side of the market is a distributor who wishes to hedge against uncertain future prices, so he purchases the futures contract on a take-or-pay basis. (With a take-or-pay contract, the distributor commits himself to purchase a given amount of electricity during specified periods on specified dates; failure to take results

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

in paying the supplier a fee.) Standardized futures contracts reduce transaction costs and are highly liquid. For example, contracts in the United Kingdom's Electricity Forward Agreement Market set quantity at a minimum of one megawatt, duration at weekdays or weekends, delivery in four-hour intervals, and settlement terms (Amundsen and Singh 1992).

Suppose today's future price for one megawatt of electricity delivered in two weeks from 2:00 pm until 6:00 pm is 10.8 cents per kilowatt-hour (kwh). Assume that, in two weeks, the spot price during that period is 10.4 cents. The distributor can either take delivery and pay the agreed upon 10.8 cents or decline delivery and pay the generator four mills per kwh for the amount of power specified in the contract. The distributor can then purchase his power requirements on the spot market at 10.4 cents. In the latter case, the generator receives four mills per kwh from the distributor who declines delivery and 10.4 cents on the spot market for its power. For both the generator and the distributor, there is no uncertainty about the price of electricity for that future period—it is 10.8 cents per kwh. The hedge is exactly the same if the spot price rose to 11.3 cents.

Futures prices are determined hourly in open auctions conducted by brokers for generators, on the one hand, and distributors and industrial users, on the other. The presence of speculators, who never take delivery on take-or-pay futures contracts, only adds bidders and sellers on both sides of the market—which serves to deepen the market and increase its liquidity and efficiency.

Electricity generators already have experience making spot market purchases of fuel, principally natural gas and petroleum, and hedging those purchases in the futures market. The combination of contracting and hedging reduces the risk born by the generator. That, in turn, reduces the risk facing the final consumer of electricity.

Similarly, distributors and large industrial users can mitigate the financial consequences of uncertain electricity price changes by entering into long-term contracts, and by hedging spot market purchases of power. Take-or-pay contracts are ideally suited for such circumstances. Futures markets already exist in Norway and the United Kingdom. New England and Mid-Atlantic power pools, in the United States, are investigating establishing futures markets.

To reduce or spread the risk associated with stochastic equipment failure or unexpected fluctuations in demand, the distributor will have signed contracts with a competitive producer carrying appropriate generating capacity. Moreover, the distributor is in a position to purchase electricity to meet service requirements on the spot market. Because the ability to obtain electricity literally on an instantaneous

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

basis is a major characteristic of service quality, it has a market price. In addition, the distributor has an incentive to discover the final consumer's demand for reliability by offering discounts for interruptible service. The added flexibility of priority service reduces the risk to the distributor that an unexpected spike in demand or equipment failure might otherwise cause.

To reiterate, the ability of a distributor to obtain electricity from a number of suppliers under long-term contract and on spot markets tends to ensure that electricity is made available at lower competitive prices, and that risk-reducing system diversity is enhanced. Intertemporal price risk arising from spot market transactions can be hedged in a futures market for electricity. At present, the development of a futures market for electric power is just beginning, but long-term contracting for electricity, spot market bids and sales, and energy wheeling are a growing part of the market for electricity in the United States.

Conclusion

The technical and economic knowledge exist to permit the substitution of market competition for state ownership or government regulation in the electricity generation industry. The chief advantages of making that substitution include a reduction of costs and lower final user prices, closer alignment between the array of services offered and consumer preferences, and greater incentives for ongoing discovery and innovation.

Generating electricity is capital intensive. Economizing on capacity dedicated to meeting peak-load demand can be a significant source of saving. Potential savings are far greater than those offered by peak-load pricing alone. Competition in generation increases load diversity, thereby reducing the inventory of standby capacity necessary to meet peak loads. The development of spot markets for electricity means that generating capacity can be more fully utilized and less idle reserves need be maintained. In addition, competition provides an incentive to discover the demand for reliability among different customer classifications. Meeting that demand more closely by offering interruptible service saves additional capital resources.

The first principle of competition is marginal cost pricing. Competitive generation provides an incentive for producers to reexamine their cost structure at every margin. Expected cost savings are greater in a competitive environment than are likely to exist under a regime of either state ownership or regulation. Under state ownership, bureaucratic self-interest runs counter to efficient, innovative operation. Regulation entails costly bureaucratic monitoring and political negotiation.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The market is an engine of discovery. Without the information about relative valuations that can only be revealed by market-determined prices, the electricity industry has no meaningful guidance for planning investment, siting plants, deciding on the array of services to offer, or reducing the effects of uncertainty. The structure and governance of the electricity industry today leads to a situation where new plants are built when demand for electricity, at politically determined prices, systematically outstrips capacity and the political will is found to proceed with construction. Siting of plants is reduced to political negotiations. Such negotiations often pay little heed to the implications of market location, fuel transportation costs, and transmission costs. Beyond time-of-day and seasonal rates, little attention is given to the quality of service. Nor is there any surprise in this. Today the industry's stated goal is to offer only the highest standard of reliability for all consumers, whether they want it or not.

Risk management in the electricity industry is synonymous with carrying enough excess capacity to meet any peak load or temporary equipment failure. Whether this approach is economic or not is never addressed. Economic valuations, reflecting consumers' choices registered in open markets, play almost no role in the decisions reached by governments, state-owned electricity facilities, or regulated utilities.

By contrast, the competitive market reduces risk by offering a choice among alternative suppliers of power, by creating spot markets, and increasing load diversification; it allows for the transfer of risk through the creation of futures markets, through prior notification of service interruption, and interruption insurance; and it spreads risk by increasing the number of producers and transactions.

The introduction of nonutility, independent power production 16 years ago in the United States, as part of a larger energy conservation policy, unleashed the forces of competition at the generation stage. The gradual spread of competitive pressures within the industry was not anticipated. Industry spokesmen and academicians have consistently missed the implications of these changes. Many have been far too cautious in estimating the efficacy of competition in enhancing the economic welfare of consumers of electrical power.[7] But perhaps this is not surprising. Market processes are discovery processes. Even informed commentators may not be very good at predicting how the market will solve many of the problems confronting an industry in transition.

[7]An influential example of such reticence is Jaskow and Schmalensee 1983. For a more recent example see Gegax and Nowotny 1993.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

References

Amundsen E.S., and Singh, B. (1992) "Developing Futures Markets for Electricity in Europe." *Energy Journal* 13: 95–112.

Baumol, W.J.; Panzar, J.C.; and Willig, R.D. (1982) *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

Berlin, E.; Cicchetti, C.J.; and Gillen, W.J. (1974) *Perspective on Power*. Cambridge, Mass.: Ballinger Publishing.

Bushnell J.B., and Oren, S.S. (1994) "Bidder Cost Revelation in Electric Power Auctions." *Journal of Regulatory Economics* 6: 5–26.

Cardos e Cunha, A. (1992) *Financial Times*, 23 January.

Caves, D.W.; Herriger, J.A.; and Kuester, K.A. (1989) "Load Shifting Under Voluntary Residential Time-of-Use Rates." *Energy Journal* 10: 83–100.

Chao, H., and Wilson, R. (1987) "Priority Service: Pricing, Investment, and Market Organization." *American Economic Review* 77: 899–916.

Christensen, L., and Green, W. (1976) "Economies of Scale in U.S. Power Generation." *Journal of Political Economy* 84: 655–76.

Coursey, D.; Isaac, R.M; and Smith, V.L. (1984) "Natural Monopoly and Contested Markets: Some Experimental Results." *Journal of Law and Economics* 27: 91–113.

Cowing, T.G., and Smith, V.K. (1978) "The Estimation of Production Technology: A Survey of Econometric Analyses of Steam-Electric Generation." *Land Economics* 54: 156–86.

Doane, M.J., and Spulber, D.F. (1994) "Open Access and the Evolution of the U.S. Spot Market for Natural Gas." *Journal of Law and Economics* 37: 477–517.

Doucet, J.A. (1994) "Coordination of Non-Utility Generation Through Priority Purchase Contracts." *Energy Journal* 15: 179–92.

Doyle, C., and Maher, M. (1992) "Common Carriage and the Pricing of Electricity Transmission." *Energy Journal* 13: 63–94.

Gegax, D., and Nowotny, K. (1993) "Competition and the Electric Utility Industry: An Evaluation." *Yale Journal of Regulation* 10: 63–88.

Goldman, O. (1923) *Financial Engineering*, 2nd ed. New York: Wiley and Sons.

Grosfeld-Nir, A., and Tishler, A. (1993) "A Stochastic Model of the Measurement of Electricity Outage Costs." *Energy Journal* 14: 157–74.

Hammond, C.H. (1986) "An Overview of Electricity Regulation." In Moorhouse, J.C. (ed.) *Electric Power: Deregulation and the Public Interest*, 31–61. San Francisco: Pacific Research Institute for Public Policy.

Hayek, F.A. (1964) *The Counter-Revolution of Science*. New York: Free Press.

Herriott, S.R. (1989a) "Competition and Cooperation in Deregulated Bulk Power Markets." In Crew, M.A. (ed.) *Deregulation and Diversification of Utilities*, 143–63. Boston, Mass.: Kluwer Academic Publishers.

Herriott, S.R. (1989b) "The Long-Run Cost Allocation Problem in the Political Economy of Electric Utility Power Pools." *Journal of Regulatory Economics* 1: 69–86.

Heyne, P. (1987) *The Economic Way of Thinking*, 5th ed. Chicago, Ill.: Science Research Associates.

Hoffman, M.C. (1994) "The Future of Electricity Provision." *Regulation* 3: 55–62.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Hogan, W.W. (1992) "Contract Networks for Electric Power Transmission." *Journal of Regulatory Economics* 4: 211–42.
Hogan, W.W. (1993) "Markets in Real Electric Networks Require Reactive Prices." *Energy Journal* 14: 171–200.
Huettner, D., and Landon, J. (1978) "Electric Utilities: Scale Economies and Diseconomies." *Southern Economic Journal* 44: 883–912.
Jaskow, P.L., and Rose, N. (1985) "The Effects of Technological Change, Experience and Environmental Regulation on Construction of Coal Burning Generating Units." *Rand Journal of Economics* 16: 1–27.
Jaskow, P.L. (1988) "Price Adjustment in Long-Term Contracts: The Case of Coal." *Journal of Law and Economics* 31: 47–84.
Jaskow, P.L., and Schmalensee, R. (1983) *Markets for Power*. Cambridge, Mass.: MIT Press.
Lyons, T.P. (1990) "Spot and Forward Markets for Natural Gas: The Effects of State Regulation." *Journal of Regulatory Economics* 2: 299–316.
Michaels, R.J. (1993) "Not Quite Free Wheeling: The Energy Policy Act of 1992." *Regulation* (Winter): 19–23.
Oren, S.S., and Doucet, J.A. (1990) "Interruption Insurance for Generation and Distribution of Electric Power." *Journal of Regulatory Economics* 2: 5–19.
Panzar, J.C., and Willig, R.D. (1977) "Free Entry and the Sustainability of Natural Monopoly." *Bell Journal of Economics* 8: 1–22.
Rose, K., and McDonald, J.F. (1991) "Economics of Electricity Self-Generation by Industrial Firms." *Energy Journal* 12: 47–66.
Rozels, R.P. (1989) "Competitive Bidding in Electric Markets: A Survey." *Energy Journal* 10: 117–38.
Schweppe, F.C.; Carmanis, M.C.; Tabors, R.D.; and Bohn, R.E. (1988) *Spot Pricing of Electricity*. Boston, Mass.: Kluwer Academic Publishers.
Sharkey, W.W. (1982) *The Theory of Natural Monopoly*. New York: Cambridge University Press.
Smith, S.A. (1989) "Efficient Menu Structure for Pricing Interruptible Electric Power Service." *Journal of Regulatory Economics* 1: 203–23.
Stewart, J. (1979) "Plant Size, Plant Factor and the Shape of the Average Cost Curve Function in Electric Power Generation: A Non-Homogeneous Capital Approach." *Bell Journal of Economics* 10: 549–65.
Strauss, T., and Oren, S.S. (1993) "Priority Pricing of Interruptible Electric Service with Early Notification." *Energy Journal* 14: 175–96.
Tschirhart, J. (1991) "Entry into the Electric Power Industry." *Journal of Regulatory Economics* 3: 27–43.
Walters, J.P., and Smith, D.W. (1993) "The Energy Policy Act of 1992—A Watershed for Competition in the Wholesale Power Market." *Yale Journal of Regulation* 10: 447–82.
Weiss, L.W. (1975) "Antitrust in the Electric Power Industry." In Phillips, A. (ed.) *Promoting Competition in Regulated Markets*, 135–73. Washington, D.C.: The Brookings Institution.
Wilson, R. (1989) "Ramsey Pricing of Priority Service." *Journal of Regulatory Economics* 1: 189–202.
Wilson, R. (1993) *Non-Linear Pricing*. New York: Oxford University Press.
Woo, C. (1990) "Efficient Electricity Pricing with Self-Rationing." *Journal of Regulatory Economics* 2: 69–81.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Sports Player Drafts and Reserve Systems

John J. Siegfried

When children play team sports they frequently "pick sides" by first identifying team captains who then alternate in selecting players. The purpose is to balance the skills of the teams so that the match is sufficiently even to interest the participants.

Limitations on the Freedom to Contract in Professional Sports

Professional sports leagues use a similar scheme to allocate players among teams. Under such a scheme the team which finishes last during the prior season has first choice (or a higher probability of receiving the first choice) from newly available players; the team with the second poorest record selects second, and so on until the championship team of the prior season chooses last. Then the process is repeated, with the weakest team getting first choice on the second round, and so on.

Such reverse order draft schemes are sometimes justified by appealing to the importance of competitive team balance among a league's members. They are effective only if coupled with a restraint which prevents a drafted player from skipping to a team which did not draft him. This complementary scheme usually consists of an agreement among the teams to refrain from employing a player drafted by another team in the league. It is called a "reserve system," after a clause inserted into baseball contracts in 1887 that once "reserved" to the team the right to unilaterally impose a new contract on a player if the team and formerly contracted player could not reach a mutual agreement for a contract extension. This clause perpetuated a team's right to a drafted player over the player's entire career.

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.
The author is Professor of Economics at Vanderbilt University.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The restriction imposed by the current draft system generally lasts for one year. A player drafted by a team for which he declines to play can reenter the draft after sitting out for a year. The player who elects to do so sacrifices a full year's earnings, which is significant for professional athletes, whose average career is less than a decade.

From time to time two leagues have operated simultaneously in the same sport in which players might participate. Under such circumstances and absent an agreement between the competing leagues, some players might find more than one team seeking their services. Alternatively, teams in different sports might compete for a player's services. Only a handful of players (e.g., Danny Ainge, Bo Jackson, Deion Sanders) are sufficiently skilled in more than one sport, however, that they might choose between the teams which drafted them in each of two different sports. Dave Winfield, for example, was drafted by four teams in three different sports when he graduated from the University of Minnesota. But, for most players the draft and concomitant agreement among teams not to tamper with their rivals' draftees gives a player only one choice: play for the team that drafted him or abandon sport as a livelihood for a year.[1]

The Effects of Monopsony

A draft and reserve system creates a situation that economists call monopsony. A monopsonist is a single buyer of an input, in this case, the labor services of a player. To maximize profits a monopsonist restricts the amount of jobs offered so that job seekers, out of fear of unemployment, become more willing to accept a lower rate of compensation. No player works for less than he could obtain in his best alternative job. Many, however, earn less of a "bonus" (or economic rent) over and above their minimum acceptable wage because they fear they will be the ones cut from the team.

Public concern about monopsony usually focuses on the exploitive salaries received by the employees. But there is also an important social welfare[2] effect of monopsony, which can be explained in the context of a player labor market. By limiting jobs to elicit restrained

[1]Playing in Canadian football, European basketball, or Japanese baseball leagues is steadily becoming a more attractive alternative than sitting out a year as salaries in foreign leagues grow.

[2]Social welfare is maximized when resources are allocated to their most valuable use. Value is assessed in terms of consumers' willingness to trade their income for goods and services. Costs reflect values that could be created if resources were devoted to their best alternative use. Social welfare is greatest when resources are employed in all those activities which command prices that exceed their costs.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

salary demands from players, teams refrain from employing some players who would create economic value (as reflected by fans's willingness and ability to pay to attend the games in which they play) that exceeds the value they can create in their next best occupation. Those marginal potential players then gravitate to their lower valued non-sports employment opportunities.

The discussion of fairness between the firms and their employees is often a peculiar one in this context, where average player compensation now exceeds $1 million in professional basketball and baseball, and sports teams are owned largely by the men and women who make up *Forbes Magazine's* annual list of the richest 400 Americans.[3] Sports-salary disputes thus hinge on judging which of two millionaires is more deserving of the net economic benefit created by the fact that sporting contests are of much more value to consumers than the next most valuable goods or services that players could create. The contrived scarcity of jobs, on the other hand, seldom makes headlines because it is difficult to see how many teams there would be, what would be their roster size, and who would be employed if there were no draft and reserve system.

Public Policy toward Contrived Monopsony in Professional Sports

The draft and associated agreement among teams not to contract with players selected by other teams is an issue for antitrust law. Section 1 of the Sherman Antitrust Act prohibits agreements among competitors which restrain trade. Certainly an agreement between Team A and Team B such that if Team A drafts Player X then Team B will not hire Player X fits this proscription.

How, then, are professional player draft schemes maintained? Major League Baseball is exempt from the Sherman Act under the 1922 Supreme Court *Federal Baseball Club* decision,[4] in which the Court determined that baseball did not meet the interstate commerce requirement for practices that can be regulated by the federal government. The Court in that case concluded that baseball is not "commerce," likening it instead to scholarly lectures. The decision has so far withstood the challenge of periodic appeals (e.g., *Flood vs. Kuhn*, 407 U.S. 258 [1971]). Recently, it has once again been under assault in Congress, this time from Senators Howard Metzenbaum of Ohio

[3]At least 24 of the "Forbes 400" wealthiest Americans in 1986 owned a professional sports team (Blitz and Siegfried 1992).

[4]*Federal Baseball Club of Baltimore, Inc. v. National League of Professional Baseball Clubs*, 259 U.S. 200 [1922].

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

and Connie Mack of Florida; but the reform bill failed to get out of committee in June 1994, and many of its supporters lost in the November 1994 congressional elections. The 1994–95 Major League Baseball Players Association strike continued to keep the issue before Congress, however.

The other significant North American professional team sports, football, basketball, and ice hockey, have avoided Sherman Act prosecution primarily through the labor exemption to the Sherman Act. This exemption relies on collective bargaining with the respective players' unions. In return for the teams freeing veteran players from the shackles of the reserve system, players' unions have agreed to continue a draft for younger players and included provision for it in their union agreement. Professional basketball is now the extreme case, where only the more promising first year (rookie) players are subject to a draft and bound to the team which drafted them (for one year). One can interpret this turn of events as a conspiracy among experienced players and team owners against new players who, of course, are not party to the union negotiations that perpetuate the draft and reserve clause for new players (White 1986).

The Effects of Player Drafts and Reserve Systems on Competitive Balance

The remainder of this article argues that (1) player draft systems (and their associated agreements against tampering with players drafted by other teams) frequently do not achieve their avowed goal of balancing competition among the teams in the league; (2) this is fortunate, because balanced competition generally would detract from society's well-being; and (3) what player draft systems really do is redistribute wealth and probably help keep some otherwise financially strapped teams in business. This view of the impotence of draft systems for promoting player skill balance across teams is the conventional wisdom among economists (Rottenberg 1956; El-Hodiri and Quirk 1971; Demsetz 1972; Quirk and Fort 1992: chap. 8) but it receives periodic challenge (Daly and Moore 1981, Daly 1992).

There are two reasons why player draft systems do not balance competition as much as owners usually claim. First, for those sports with a large number of players per team (e.g., baseball and football) one good prospective player per year does not make much difference in overall team talent. But that is the difference a draft makes—one good prospective player per year. The team with the poorest record in the prior year chooses first. At the end of the first round of player selection the prior year's champion selects, followed directly by the

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

second pick of the team with the poorest record. Except for the beginning of the first round, the championship team and the worst team in the league receive consecutive choices. If there is a continuum of player skills and uncertainty about how new players will perform at the professional level, the team picking second among consecutive selectors can easily get the better players. That reduces the advantage of the weaker team to only its first-round selection. In basketball and ice hockey one excellent player sometimes makes a difference, but for football and baseball the advantages given to weaker teams by the draft system work slowly, if at all.

The second reason player draft systems do not balance competition is captured in the well-known Coase Theorem which states that if property rights are well defined and the costs of buying and selling those rights are not prohibitive, resources will eventually gravitate to their most valuable use no matter who is given the rights initially (Coase 1960, Demsetz 1972). If the rights are given to the party that values them the most, no one else will make an offer for those rights that exceeds their value to the initial owner, and no transfer will occur. If, on the other hand, the rights are initially assigned to a team that values them less than does another, an opportunity to trade those rights exists in which both the initial owner and the purchaser can end up better off. Consequently, the rights would be sold to the team that values them most.

Thus, if a highly skilled new player is worth more to, say, the prior year's championship team, located in a large metropolitan area, but is drafted by the weakest team in the league, located in a low-population area, the weak team might be better off selling its rights to acquire the new player to last year's champion. The teams could agree on a price that exceeds the added revenues the weak team would enjoy if it retained the new player, but is less than the added revenues the new player can stimulate in the championship city. The result is that relatively weak teams draft the top new players, and promptly sell them to their more successful rivals. Since the sale is voluntary, both teams must be better off, including the cellar-dweller that continues to lose. This can explain episodes of talented teams selling off players for cash—for example, the 1960s Kansas City Athletics, the 1970s Oakland Athletics, or the 1993 San Diego Padres.

There may be some limits to such player transfers beyond the requirement that property rights to the players' services be well defined, and transactions costs be low. For example, teams in a league presumably will not continue transfers to the point where both lose revenue because league competition gets so lopsided that fans begin to find the games uninteresting and switch their allegiance to opera

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

or bird-watching. And transfers during a season may create such confusion among fans that interest wanes sufficiently to make "no-transfer" periods profitable. In general, however, because of trades and sales, player drafts do not lead to the level of playing-skill balance that team owners sometimes claim they achieve.[5]

The Welfare Effects of Balanced Competition

For the sake of social welfare, this failure to balance competition is fortunate. The argument that balanced playing competition is essential to a league's survival is usually founded on the assumption that consumers value only one characteristic of sporting contests, namely uncertainty of outcome. While there is substantial empirical evidence that attendance is higher when evenly balanced teams compete (Knowles, Sherony, and Haupert 1992), there is also strong evidence that attendance responds to winning, and the response varies by team (Noll 1974, Whitney 1988, Scully 1989, Porter 1992). Consumers are willing and able to pay more for a winning team in some places than in others. Preferences are not identical. For example, on average Canadians may be willing to pay more for a winning ice hockey team than the typical U.S. resident (Noll 1974).

Even if average consumers value winning similarly in all locations, the rewards for winning still will vary by location. An identical propensity of individuals to attend a sporting contest in two cities will result in similar total attendance only if the two cities have a similar population. It is total revenue, not revenue per consumer, that team owners bank, and there will be more ticket revenue for the owner of a winning team in a larger-population city than in a smaller-population city.[6] Even though an imbalance in team-playing skills will reduce the uncertainty of games and thereby the interest in (and attendance at) those games, the greater aggregate willingness to pay for winning by certain teams than for others can mean that the profit-maximizing degree of competitive balance is far from a perfectly even balance. In a simulation model of a professional baseball league, using actual data from Major League Baseball, Joseph Hunt and Kenneth Lewis (1976) discovered that the equilibrium dominance of the most success-

[5]If the draft and a reserve system promote team balance, then the introduction of free agency in baseball in 1976 should have been accompanied by increased competitive team imbalance. The experience in baseball since 1976 reveals exactly the opposite, however. There has been more competitive balance since free agency arrived than during the period of restricted labor markets (Zimbalist 1992: 95–101).

[6]Local media revenues also are likely to vary with population, exacerbating differences in potential revenue from large and small markets.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

ful teams in a six-team baseball league had not been achieved even by the dominating New York Yankees of the 1950s and 1960s.

For the sake of economic efficiency, it is fortunate that player draft systems do not achieve the goals they are allegedly set up to achieve—competitive balance. Imbalance leads to more winning in those places where people get more satisfaction from winning, and less winning where people in aggregate do not care as much about it.

The Real Purpose of Labor-Market Restrictions in Professional Sports

If player drafts do not achieve competitive balance, and if competitive balance would reduce both teams' profits and social welfare in any case, why then do leagues maintain such an interest in player draft and reserve systems? The answer, perhaps, lies in the wealth-distribution effects of a player draft (White 1986; Quirk and Fort 1992: 235-39). A draft places certain players (namely rookies) in a weaker bargaining position than they would otherwise enjoy.[7] This means that those players subject to the draft will be paid less than they would otherwise earn in an open market.[8]

There is empirical support for this proposition. Prior to the 1970s players largely were bound to the original team that drafted them (or another to which the original team had transferred their rights) for life. When unions formed in the various sports, and collective bargaining began, veteran players in all of the leagues negotiated some relief from this restriction. Professional ice hockey and football players enjoyed the least freedom to negotiate with other teams (until 1993 for football). Baseball players are freed completely from the reserve clause only after six years experience in the major leagues.[9] Basketball players get relief from monopsony exploitation after just one year in the National Basketball Association. Statistical studies of salaries show convincingly that when draft and player-retention schemes are relaxed, large increases in player compensation result (Sommers and Quinton 1982; Raimondo 1983; Hill 1985; Scott, Long, and Scomppi 1985;

[7]New players' bargaining position, particularly for a long-term contract, may remain strong in an absolute sense, however.

[8]Quirk and Fort (1992: 238) conclude that salaries in professional baseball became noticeably less equally distributed as free agency was ushered in after 1976. They find that a disproportionate share of the benefits of free agency accrued to experienced players.

[9]Mandatory arbitration of salary disputes for players with three to five years experience in major league baseball cause some of the effects of free agency to migrate to players with fewer than six years experience because arbitrators can base their decisions on the salaries of comparable players with an additional year of experience vis-à-vis the player in arbitration.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Scully 1989; Kahn 1993; Quirk and Fort 1992). Compensation grew most rapidly at the times the draft and reserve schemes were constrained by either the courts or collective bargaining in each sport, and have grown faster the more relief from the draft and reserve system has been won by the players. Over the past 20 years professional basketball players' salaries have risen fastest, baseball players' second fastest, and football players' slowest (Staudohar 1989). The rapid increase in football player salaries in 1993, the first year of true free agency for veteran players, further corroborates the connection between salary levels and relaxed monopsony restrictions.

The redistribution of wealth resulting from a player draft is not only between team owners and players, and between veterans and rookies. There is also a redistribution of wealth among team owners as those who draft players whose value is greater elsewhere capture some of the differential value through the transfer price. Consequently, generally more successful (winning) teams redistribute wealth to the perennial losers. This helps some financially marginal teams survive, and, unfortunately, means that teams endure in certain cities where fans are not sufficiently interested to pay more than the value of the players' time devoted to some other endeavor including playing that sport in another city.

Conclusion

The extent to which a player draft and reserve system affects social welfare depends on the ease with which players' contracts can be transferred among teams. The greater the freedom of teams to buy and sell players' contracts, the less balance will be created by a draft and reserve system, and the less damage it will cause to social welfare by shifting winning from places where consumers value it more to places where consumers care less about "being number 1."

If player contracts cannot be transferred easily, then labor market restrictions that convey monopsony power on teams will favor consumers' preferences for uncertain outcomes of contests to the exclusion of differences among consumers in preferences for winning. Consumer welfare is likely to suffer because the efficient balance between winning and contest uncertainty will be missed. If, on the other hand, player contracts can be transferred easily among teams, then the effects on social welfare of a draft and reserve system are innocuous. Under such circumstances, the primary effects of monopsony power in professional sports are on the distribution of wealth among teams, between team owners and players, and among the players themselves.

Social welfare could also be improved on the playgrounds of America if, after choosing sides for a "pick-up" game based on a mechanism

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

designed to balance team strength, those players who value winning more could pay off those who value winning less for the right to rearrange the teams more to the satisfaction of those who think winning is the only thing.

References

Blitz, R., and Siegfried, J.J. (1992) "How Did the Wealthiest Americans Get So Rich?" *Quarterly Review of Economics and Finance* 32: 5–26.

Coase, R.H. (1960) "The Problem of Social Cost." *Journal of Law and Economics* 3: 1–44.

Daly, G. (1992) "The Baseball Player's Labor Market Revisited." In Sommers, P.M. (ed.) *Diamonds Are Forever: The Business of Baseball*. Washington, D.C.: The Brookings Institution.

Daly, G., and Moore, W.J. (1981) "Externalities, Property, Rights, and the Allocation of Resources in Major League Baseball." *Economic Inquiry* 19: 77–95.

Demsetz, H. (1972) "When Does the Rule of Liability Matter?" *Journal of Legal Studies* 1: 13–28.

El-Hodiri, M., and Quirk, J. (1971) "An Economic Model of a Professional Sports League." *Journal of Political Economy* 79: 1302–19.

Hill, J.R. (1985) "The Threat of Free Agency and Exploitation in Professional Baseball: 1976–1979." *Quarterly Review of Economics and Business* 25: 68–82.

Hunt, J., and Lewis, K. (1976) "Dominance, Recontracting and the Reserve Clause: Major League Baseball." *American Economic Review* 66: 936–43.

Kahn, L.M. (1993) "Free Agency, Long-Term Contracts and Compensation in Major League Baseball: Estimates from Panel Data." *Review of Economics and Statistics* 75: 157–64.

Knowles, G.; Sherony, K.; and Haupert, M. (1992) "The Demand for Major League Baseball: A Test of the Uncertainty of Outcome Hypothesis." *The American Economist* 36: 72–80.

Noll, R.G. (1974) "Attendance and Price Setting." In Noll, R.G. (ed.) *Government and the Sports Business*, chap. 4. Washington, D.C.: The Brookings Institution.

Porter, P.K. (1992) "The Role of the Fan in Professional Baseball." In Sommers, P.M. (ed.) *Diamonds Are Forever: The Business of Baseball*, chap. 4.

Quirk, J., and Fort, R.D. (1992) *Pay Dirt: The Business of Professional Team Sports*. Princeton, N.J.: Princeton University Press.

Raimondo, H.J. (1983) "Free Agents' Impact on the Labor Market for Baseball Players." *Journal of Labor Research* 4: 183–93.

Rottenberg, S. (1956) "The Baseball Players' Labor Market." *Journal of Political Economy* 64: 242–58.

Scott, F.A.; Long, J.E.; and Scomppi, K. (1985) "Salary vs. Marginal Revenue Product under Monopsony and Competition—The Case of Professional Basketball." *Atlantic Economic Journal* 13: 50–59.

Scully, G.W. (1989) *The Business of Major League Baseball*. Chicago: University of Chicago Press.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Sommers, P.M., and Quinton, N. (1982) "Pay and Performance in Major League Baseball: The Case of the First Family of Free Agents." *Journal of Human Resources* 17: 426–36.

Staudohar, P.D. (1989) *The Sports Industry and Collective Bargaining*. Ithaca, N.Y.: ILR Press.

White, M.D. (1986) "Self-Interest Redistribution and the National Football League Players Association." *Economic Inquiry* 24: 669–80.

Whitney, J.D. (1988) "Winning Games vs. Winning Championships: The Economics of Fan Interest and Team Performance." *Economic Inquiry* 26: 703–24.

Zimbalist, A. (1992) *Baseball and Billions*. New York: Basic Books.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The Performance and Stability of Banking Systems under "Self-Regulation": Theory and Evidence

Michael D. Bordo and Anna J. Schwartz

Recent literature that makes the case for the elimination of government regulation of a banking system rests on a variety of theoretical propositions. Each of these propositions, addressing the question whether self-regulation by banks will set an external limit on their issues, reaches an affirmative conclusion. We propose to test the cogency of three of these propositions and to examine the validity of the empirical evidence that has been adduced as confirmation of the theory. The three propositions are:

1. The operation of an interbank clearing mechanism checks overissue by individual banks and, in some versions of the theory, by the system as a whole.
2. Private issuers in a competitive system have incentives to establish confidence in the value of the moneys they produce and therefore limit the quantities they can safely emit.
3. Restriction of discounting by banks to real bills limits the quantity of money they issue.

Those propositions do not define whether government retains a role as an issuer of outside money. They relate only to whether government should intervene in the provision of inside money by private sector banking institutions. The empirical evidence that bears on the propositions we here consider is drawn from monetary systems in which banks were regulated in a variety of ways and in which outside money of specie was supplemented by government-controlled fiat-money elements. The performance of the banks in those systems

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

Michael D. Bordo is Professor of Economics at Rutgers University and a Research Associate of the National Bureau of Economic Research. Anna J. Schwartz is a Research Associate of the National Bureau of Economic Research.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

cannot be evaluated as reflecting solely the particular theoretical proposition the evidence is supposed to support. The performance is obviously also related to the existing regulations and inside money-convertibility requirement. It is therefore not easy to sort out the contribution of the theoretical proposition.

In any event, proposing a rationale for eliminating government intervention in the provision of inside money does not adequately define a monetary system. It is essential in addition to discuss the conditions for the provision of outside money. In our view, a more fundamental question than whether government intervention in the provision of inside money is desirable is whether the institutional arrangements for the provision of outside money produce a relatively stable monetary environment. A prerequisite of good performance by a banking system is such an environment that only the outside money arrangements can assure.

The paper proceeds as follows. We begin by discussing inside money in the first three sections, which in turn present each of the theoretical propositions and the supporting evidence. We find the theories to be flawed and the evidence unconvincing. In the fourth section, we suggest an alternative approach to "self-regulation" that would free banks to establish offices where they chose, to pay whatever interest was required to obtain funds, and to acquire assets yielding the highest return as they judged it. The alternative approach is based on two principles: first, the system is so designed that risk is borne by owners and managers, not by money holders and taxpayers; second, the provision of outside money produces a stable monetary environment. In a stable monetary environment, price level stability prevails. As a result, if banks make mistakes in acquiring assets, the mistakes are attributable to faulty credit analysis, not to price level or inflation surprises. Available historical evidence can be cited in support of the approach only in the negative sense that the evidence violates one or the other principle. In the fifth section, we discuss the current worldwide fiat-outside-money regimes, and ask whether they can be constrained to limit the outside money they supply. We take for granted that outside-money arrangements will remain in the hands of governments. Provision of outside money by government raises the question whether it should exercise the role of lender of last resort. We do not, however, deal with that question. The final section summarizes our conclusions.

The Clearing Mechanism

Theory

Adam Smith was among the first to propound the doctrine that a clearinghouse provides an automatic mechanism of adverse clearings

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

that serves to regulate the issues of its members. Originally the doctrine applied to note issues, but later it applied to the clearing of both notes and deposits. The doctrine is a prominent feature of recent advocacy of a self-regulated banking system.

The essentials of the process by which a deposit in one bank leads to an expansion of the aggregate money supply as a multiple of the reserves that banks maintain were also understood early on and have been restated in the recent literature. Under the clearing mechanism, a bank that increased its issues disproportionately to those of other banks would experience a drain of its reserves. It would find balances running against it at the clearinghouse and be forced to contract. The mechanism thus provides a check to individual bank overissue. Advocates stress the importance of frequent clearing of notes and checks as means to prevent overissue. In the main, however, neither the early writers on the subject nor more recent ones have understood that a clearinghouse provides no check if all banks expand simultaneously.

The theory of the banking firm underlying the clearinghouse paradigm is based on a set of fixed coefficients, of which the reserve ratio is the one most often cited. For a dollar of additional deposits in a particular bank, the fixed reserve multiplier defines the extent to which the aggregate money supply increases. A set of fixed coefficients, however, does not illuminate the role of market conditions like interest rates and uncertainty that affects the holding of reserves. The theory of clearinghouses as regulators of bank issues is thus derived from a rudimentary level of analysis.

The chief error of the clearinghouse paradigm is to infer that the limits to expansion to which an individual bank is subject apply also to the banking system as a whole acting in unison. Lawrence H. White (1984: 17) does not make this error, noting that adverse clearings will not arise among a group of banks sharing a common expansion. George Selgin (1988: 80) attempts to respond to this criticism of the effectiveness of a clearing arrangement as a method of controlling the supply of money by arguing that "spontaneous in-concert expansions will be self-correcting" because the growth in total clearings will increase the variance of clearing debits and credits.[1] Banks in his view will protect themselves against the risk of default at the clearinghouse at any clearing session, leading to a unique equilibrium supply of inside money.

[1]Yet in discussing Australian experience, Selgin (1988: 50) notes that "prices were fairly stable, and the principle of adverse clearings insured that no single bank could step out of line with its competitors. If by chance the entire system went out of line, adjustment would come as a consequence of gold losses abroad." He here ignores his own dictum of self-correcting in-concert expansions.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Clearinghouses were formed to minimize the costs of note exchange and check clearing but in time acquired a monitoring and supervisory role in the banking industry. The clearinghouse in that aspect provided self-regulation for the industry. Membership in a clearinghouse attested to the adequacy of a bank's capital. Audits by the clearinghouse which also required the banks to publish statements of condition supplemented the regulatory demands of government supervisors. Most of the recent discussion of the role of clearinghouses does not emphasize their direct supervisory activities in enforcing limits on overissue by member banks and thus creating monetary confidence but rather concentrates on the restraint imposed by adverse clearings. One commentary, however, links clearinghouse monitoring of the product quality of demand deposits that served also to control the behavior of bank managers to the information-related disadvantages of checks (Gorton and Mullineaux 1987: 458–60).

Evidence

The principle of clearing means of payment was known and employed even in ancient times but the modern clearinghouse dates from 1760 in Edinburgh, followed by London in 1773, and decades later in other European cities. In the United States, the first clearinghouse was established in New York in 1853 and in five other cities before the outbreak of the Civil War.

The evidence most often cited in support of the significance of clearing and redemption systems in controlling the aggregate money supply includes the performance of the Scottish banking system before 1845, the First and Second Banks of the United States, the Suffolk Bank in Massachusetts, and Canadian clearing arrangements (Dowd 1989). We comment on each of these exemplars in turn.

Scottish Banks. Except for three chartered institutions, Scottish banks from 1749 operated as partnerships, under individual ownership or as unregistered joint stock companies, enjoying freedom of note issue (except for the prohibition of small notes after 1829) with the owners subject to unlimited liability for the institutions' debts. The Bank Act of 1844 in England ended freedom of issue in Scotland for all but the nine banks in existence at that date. A new bank would be denied the right of note issue, but an existing note-issuing bank could exceed its authorized circulation if collateralized by 100 percent marginal specie reserves.

The operation of the Scottish system has been adjudged a success relative to that of the English banking system, where the Bank of England possessed certain monopoly privileges, and limits on the size of other banks were legislated.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

In accounting for the "outstanding success of the Scottish free-banking system," Roland Vaubel (1984: 61), a sympathetic commentator, mentions freedom of entry, absence of limited liability, and no limits on the size of banks. He does not refer to the operation of the clearing system, a significant omission that implies it had no importance. White, another sympathetic commentator, treats the clearing arrangements as "self-policing" in that clearing losses restrain an individual bank from expanding faster than the average. In discussing the restraint on the banking system as a whole, however, he stresses not the clearinghouse but the role of internal and external reserve drains that guaranteed convertibility of notes into specie effected (White 1984: 1–22). The Scottish experience therefore is inconclusive on the role of the clearing arrangements in limiting expansion by the system as a whole.

First and Second Banks of the United States. Both the First Bank (1791–1811) and the Second Bank (1816–36) promptly presented state bank notes for redemption and served to that extent as clearinghouses that regulated the currency, although in their own operations their performance was not invariably above reproach. Statistics on the operation of the First Bank do not exist, but it apparently extended permanent loans to individuals and banks (Holdsworth 1910: 124).

The Second Bank was mismanaged before 1819 when Western and Southern branches overissued their notes, which were redeemable in the East, until redeemability of all notes at all branches was discontinued (Schweikart 1991: 607–18). After that date, under Nicholas Biddle's tutelage, the Second Bank created a uniform currency by forcing the redemption of state bank notes (Fraas 1974). It then expanded rapidly in 1828–31 and subsequently, with Jackson's veto of its charter and the removal of Treasury deposits, it was forced to contract.

The Suffolk Bank. Founded in Boston in 1818 as an ordinary bank of issue and deposit, its management decided a year later to redeem country bank notes it purchased at a discount from merchants, individuals, and other banks. It allowed the issuing banks to purchase their notes from the Suffolk at the same discount, provided they made permanent deposits in the Suffolk of $5,000 plus additional deposits to cover notes redeemed. Discounted notes of a non-participating bank were immediately returned for redemption at par. The arrangement proved to be unprofitable for the Suffolk since apparently investment of the permanent deposits provided the only source of net income.

In 1825 the bank modified the arrangement by establishing an interest-free fund of $300,000 assessed on the Associated Banks in

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

proportion to their capital, paid in the notes of the participants, with the Suffolk as their agent. The aim was to reduce the size of the discount on notes. Within a year, notes of the banks in good standing were accepted at par at the Suffolk. Individual bank behavior was subject to control by the Suffolk, which could suspend or expel a bank if its permanent deposit were eroded by excessive note issues (Mullineaux 1987: 888–921).

We observe that although the Suffolk bank unquestionably provided a check on overissue by an individual bank, the evidence of its total annual redemptions from 1834 to 1857 contradicts the theory of "self-correcting in-concert expansions." Not only is there a clear upward trend in the figures, rising from $76 million in 1834 to $376 million in 1857, in line with a growing banking industry, but in addition every cyclical upturn and downturn in the period is registered in the time series (Dewey 1910: 89). The Suffolk system was no bulwark against instability in the monetary system.

Canadian Clearing Arrangement. After confederation in 1867, Canada permitted branch banking and authorized chartered banks to issue notes equal to their paid-in capital, but required them by law to establish redemption agencies at the main city in each of the then seven provinces and elsewhere as determined by the Department of Finance. The chartered banks exchanged notes daily at clearinghouses in each of the larger cities and their branches at other centers. Balances were paid in Dominion notes, a government-issued currency, or in drafts on commercial centers. Chartered bank notes circulated at par. Legal tender was either Dominion notes or gold. Some bank failures occurred, but in 1907, unlike the panic in the United States, no restriction of payments took place in Canada, despite a currency drain. For that reason, in 1908, the limit on note issue was raised to 115 percent of paid-in capital during the crop moving season (Schuler 1992: 88).

It is not obvious, however, that the clearing arrangement rather than branch banking, competitive note issue, or convertibility contributed most to the good record of Canadian banks before 1914. After that year the Department of Finance acted as a quasi-central bank, and in 1935, the Bank of Canada was established (Bordo and Redish 1987: 402–06).

Private Competitive Note Issue

Theory

This is an argument against the classical view that unregulated competition in the supply of fiduciary money would lead to overissue,

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the destruction of its value, and an infinite price level. True, as argued in the classical view, if different private money producers offered indistinguishable homogeneous monies, convertible into one another at fixed exchange rates, overissue would be possible. For competitive production of fiduciary money, however, according to the recent view, output differentiated by brand names is essential. By relying on the brand name attached to each producer's money, consumers can acquire information about the quality of the monetary service flow from each one. The brand name is not only information for the consumer but capital for the producer (Klein 1974: 431–35).

When monies are differentiated by issuer, and each is valued relative to competing monies at flexible exchange rates, overissue by a single producer might gain him a short-term profit, but at a cost of disinvestment in the producer's brand-name capital. A decision by a producer of money to overissue would be at the expense of his long-run profit in making his product acceptable to the market. A competitive money producer creates confidence in his product by limiting its issue.

Flexible exchange rates among private competing monies may be a condition of a system with fiat outside money, not of a system with commodity-based outside money. If one assumes the latter, in a contemporary setting, would competitive bank notes trade at discounts in cities distant from the place of issue, given low modern communications and transportation costs? The conclusion reached by Lawrence H. White (1989: 372) is that nowadays, in a competitive system with unrestricted branch banking, all banks would be driven to accept one another's liabilities at par. They would do so either to profit from replacing rivals' notes with their own or to enhance the demand for liabilities of all banks thanks to par acceptance. The check on overissue would be the adverse clearing mechanism.

Richard Cooper (1989: 393–94), however, finds this scenario implausible. He asks whether government would not inevitably intrude in a hypothetical private competitive banking system if only to decide which bank notes it would accept in payment for taxes, and which it would use in making its own payments. If the government set high standards for banks whose notes it would accept or use, ease of entry and free banking would become limited.

Evidence

Although selected episodes of historical experience have been interpreted by some observers as representing flexible exchange rates between competing monies, the author of the theory (Klein 1974) finds that conclusion dubious when he discusses historical examples. The reason is that in these examples competing monies were all

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

convertible on demand into a single outside money that served as the unit of account and in fact were closer to internal fixed exchange rate monetary arrangements than to internal flexible rates. The historical evidence suggests that such arrangements arose to minimize money changing and valuation costs (that is, transaction and information costs) of competitive independent monies.

In a recent study Kurt Schuler (1992) identifies about 60 historical cases of what he defines as competing free banks, which lasted from a few years to over a century. Free-banking systems included many that were regulated and a dozen or so that were not. The absence of a central bank was the critical factor in determining whether a banking system qualified as one with competing free banks. Many such cases were colonial possessions of European imperial powers. The study does not examine whether flexible exchange rates between the competing monies were a feature of the historical cases.

The literature on competing banks focuses on their performance as note issuers, in part reflecting the historical period when bank liabilities were predominantly notes. The main reason for this focus, however, is the emphasis in the literature on the malign consequences of the monopoly of note issue by central banks. The examples are intended to convince readers that the record of competing banks as note issuers was beneficent and that competitive banking systems without central banks in today's economies would be superior to existing monetary arrangements.

We report examples of competitive monies that are commonly cited, beginning with evidence for colonial America and the United States, then for the United Kingdom and Europe, and for East Asia and Oceania. The questions that need to be answered include the following: Were the banks self-regulated? Were they competitive? Was overissue avoided? Did macroeconomic stability prevail?

The United States

New England Colonial Issues. Bills of credit issued by the Massachusetts, Connecticut, Rhode Island, and New Hampshire colonies in the first half of the 18th century were accepted at par under an arrangement with the other New England colonies in payment of taxes and in general exchange, stimulating overissue. As a result, the paper monies of the New England colonies experienced a greater depreciation than those of the middle colonies, where issues of different colonies (in the absence of the type of arrangement that prevailed in New England) circulated domestically at flexible exchange rates. New England government issues with fixed exchange rates among them is consistent with the theory of the conditions for their overissue.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Merchants petitioned Parliament against the colonial government bills of credit. Their issue in the four New England colonies was prohibited in 1751.

Various land bank schemes were also projected in the colonies, some under private, and others under government sponsorship. The land bank private experiments survived at most for a year, suppressed either by the colonial or British legislature. Public land banks were organized in 12 of the 13 colonies, issuing notes on the security of land. When the size of the loans made was limited, and their term fixed, overissue was not a problem—the famous example of Rhode Island to the contrary notwithstanding.

State Chartered Banking, 1790–1816. Competitive chartered banks existed in the states, issuing their own notes. Their unimpressive record has been analyzed as a failure, not of private currency competition, but of state interference with the banks' production and pricing decisions, "since they [the states] had a financial interest in their expansion even at the price of inflation and irredeemability" (Vaubel 1984: 67). The states are also held responsible for their failure "to provide an adequate legal framework for the competitive part of the process: to ensure information, to enforce contracts, and to prevent or prosecute fraud." These criticisms seem to imply a need for regulation of banks, reinforced by the critic's statements of approval for annual examinations of banks by state-appointed regulators.

Free-Banking Era. So-called free-banking laws in New York (1838) and Michigan (1837) permitted free entry—a requirement for a competitive system—but enforced capital requirements, specie-reserve requirements, and state bonds as security for note issues—forms of regulation. Sixteen other states followed suit, but in only eight were free banks established. Distinguishable bank notes circulated at varying discounts from par determined by their distance from cities where they were presented for redemption. Merchants and brokers accepted the notes at these discounts.

The literature on free banks before World War II portrayed them as wildcatters. Recent studies find the earlier reports of wildcatting exaggerated, the losses noteholders experienced minimal, and contagious runs on banks limited in extent (Rockoff 1975; Rolnick and Weber 1983: 1080–91; 1984: 267–91; Schweikart 1991: 621–23). Some state free banks were successful; many were not, and failure rates were considerably higher in free bank than in non-free bank states (Kahn 1985: 883). Free banks failed either because of capital losses on the bonds in their portfolios that depreciated with unstable state finances, possibly because the banks' performance was imprudent (Rockoff 1991: 73–109).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Yet the demise of state free banking had nothing to do with a decline in market demand for those banks. The end came because of the enactment of a national banking system authorized to issue notes backed by federal bonds, and of a prohibitive tax, effective July 1866, on state bank notes.

Because most state bank notes did not fluctuate widely in terms of one another and in terms of specie, and because inconvertible notes, no matter how large the discount, were unacceptable in transactions, one commentator (Klein 1974: 440) reached a skeptical conclusion on the significance of this episode: "Monetary arrangements during the 19th century free-banking era were much closer to multiple monies circulating at fixed exchange rates than to multiple monies circulating at flexible rates."

A different question concerns the assumption that free banks lowered barriers to entry in the U.S. banking industry. The number of banks after their introduction in states with free-banking laws was either unchanged or declined (Ng 1988: 877–89). Hence the competitive banking paradigm lacks confirmation on this score as well as with respect to the observation of flexible exchange rates among competing monies.

In any event, the free-banking era did not coincide with economic stability in the United States. It coincided with the disturbance to the monetary system produced by the great gold discoveries in California in 1848 and subsequently in Eastern Australia and elsewhere (Rockoff 1991: 77). Even so, the bimetallic standard then prevailing might have produced a stabler and slower rate of monetary growth than would have been the case under a fiat-money system. Whether the historical record of state free banks lends support to the case for the adoption of free banking now is not easy to sort out.

The United Kingdom and Europe

Free Banking in Scotland before 1845. The Scottish competitive note issue has been discussed above to assess the contribution that clearing arrangements made to what has been described as a century of stable free banking that ended with the Bank Charter Act of 1844 and the Act of 1845, which, as noted, restricted Scottish note issue. Here we examine the episode to determine if it conforms to an essentially unregulated free-market monetary system. The episode clearly does not conform to the theory of flexible exchange rates among distinctive monies, since Scottish bank notes circulated at par, with no variations in their market value reflecting differences in perceived qualities.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Other features of the Scottish system have also been interpreted as deviations from a free-market system. Limited liability charters were restricted to three banks, and all other banks operated with unlimited liability. The impact of unlimited liability on free banking as a competitive system is an unresolved issue. Unlimited liability may be a barrier to entry and a limit on the size of firms, since access to organized capital markets is a benefit of limited liability (Carr and Mathewson 1988: 766–84). However, the chartered banks had no competitive edge in note issue or intermediation, and note exchange among the provincial banks emerged before the chartered banks decided to join the system (L.H. White 1991: 19).

Even the role of the note exchange system has aroused controversy, since free-bank advocates imply that all note issuers benefit from it and should therefore have voluntarily joined. But some did not, and the larger banks exercised market power to compel their membership (Munn 1985: 341–43).

Another subject of debate has been the interpretation of the comparative failure rate of Scottish and English banks. Seventeen banks went out of business during the 100 years of the Scottish free-banking period ending in 1844, a failure rate one-quarter that of English banks, and losses to note holders and depositors were less than half that of banks in the London area alone. For subsets of the century, however, the Scottish failure rate was not invariably lower than the English failure rate (Sechrest 1988: 251–52). Low failure rates themselves have been challenged as unacceptable indicators of superior performance (Rothbard 1988: 230). The prohibition of adequately capitalized and diversified joint stock banks presumably was responsible for the English system's instability. Was the failure rate of Scottish free banks with no limits on the number of partners proof of greater stability of free banks than regulated banks, or merely a reflection of the importance of adequate capitalization?

Also in dispute is the possible dependence of the Scottish banks on the London money market. Each of the banks is said to have held its own specie reserves, but, as has been suggested, in periods of stringency did the chartered banks act as lenders of last resort to the provincial banks and did the chartered banks in turn look to the London money market and the Bank of England for support? (Goodhart 1987: 129–31). Cases of borrowing from the Bank of England by Scottish provincial and chartered banks do not, however, appear in bank archives (Munn 1991: 64).

Finally, doubts are raised about claims for the superior performance of Scottish banks by instances of absence of full convertibility of their note issues, even after the Act of 1765 denied the legality of option

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

clauses and affirmed redeemability on demand. One commentator regards imperfect convertibility as limiting the banks' costs and dangers and thus promoting the success of Scottish banking (Checkland 1975: 186). Alternatively, deferring redemption until a future date may be indicative of the system's weakness.

The upshot of the debate on the Scottish banks is that the extent to which they were a model of a private, competitive, unregulated, and independent system remains unsettled. Convertibility of notes into specie, the unlimited liability of the owners of banks, who became personally liable for their issues, and the familiarity of note holders with the private means and standing in the community of the owners seem to have been factors accounting for the relative stability of the Scottish competitive banks before 1845.

Free Banking in France, 1790–92, and 1797–1806. Two experiments in free banking occurred in France. The first was a satellite of government-issued assignats, printed only in large denominations and that drove coin from circulation. So-called *caisses patriotiques* were established by towns, merchants, manufacturers as well as bankers to exchange assignats for small-denomination notes. Originally 100 percent reserve banks that charged a small transaction fee, in time they became fractional reserve banks that operated without government regulation, engaging in discounting commercial bills. Charges of counterfeiting and excessive issues of the notes—known as *billets de confiance*—have not been substantiated, but attacks in the press, the refusal of some government tax collectors to accept the notes, and a National Assembly proposal that the caisses deposit their reserves with local authorities undermined the confidence of the note holders, who staged a run on the banks. On November 8, 1792, the Assembly voted to close the caisses by the end of the year. Few needed government help to redeem their notes. Although inflation was generated not by overissue of billets de confiance but of assignats, the revolutionaries adopted price controls to contain it, closed the bourse, and prohibited trade in specie. By mid-1793, the Reign of Terror ended most private banking (E. White 1991: 131–39).

The revolution's original espousal of free enterprise was revived by the Directory, which repealed anti-banking laws in 1796. New banks of issue were soon established, the structure of the system much like that of the caisses patriotiques—two or three dominant banks in Paris, accounting for a large fraction of all notes issued, and other relatively small banks both in Paris and the provinces. The large banks redeemed their high denomination notes in silver or gold coin, while the small banks redeemed their small note issues in copper coin. As in the U.S. free-banking era, issues of the small banks circulated at small

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

discounts. Competition prevailed, with no widespread failures or counterfeiting. A defalcation at a large bank in October 1798 threatened its solvency, but the Treasury offered a loan, shareholders deposited coin to guarantee the bank's liabilities, and a run was avoided.

Shortly after Napoleon seized power in 1799, he established a bank to serve his purposes, the Banque de France. The large banks in Paris were forced to convert their shares to shares of the Banque and the small banks were dissolved; in 1803 the Banque obtained the exclusive privilege of note issue for 15 years in Paris. The circulation of the provincial banks was relatively slight. At Napoleon's insistence the bimetallic standard was reestablished in 1803, which constrained the revenue the government could raise from money creation. In 1806 he appointed his choices for governor and deputies, increased the bank's capital, and extended to 25 years its monopoly of issue in Paris. It lost its monopoly from 1815 to 1848 and had to compete with private banks of issue in the provinces. These were regulated banks whose notes were restricted to the provinces in which they operated. In 1848, however, the Banque de France became the exclusive issuer of notes in France (E. White 1991: 139–44).

Competitive Banking in Switzerland, 1825–91. Until the National Bank of Switzerland was established in 1905, after 25 years of debate, and given a monopoly of note issue, as many as 36 provincial state-incorporated banks and private bankers had engaged in the note-issuing business—the first one in 1826, according to one source, in 1834, according to another, who does not refer to an earlier cantonal bank (Weber 1988: 461–62; Landmann 1910: 9). The notes initially circulated only locally at varying discounts. Few currency issuers defaulted before 1850. Freedom of entry existed even after uniform reserve requirements were imposed in 1881. Notes thereafter were limited to double the amount of share capital, and banks were required to accept in payment at full value their own as well as other banks' notes (Landmann 1910: 9–63, 204–38). The Swiss banks, it has been argued, depended on Paris in times of stringency, much as the Scottish banks before 1845 are said to have depended on the London money market and the Bank of England (Goodhart 1987: 131).

Competitive Banking in Sweden before 1897. The Riksbank, in existence since 1656, remained as the only bank in Sweden in the 1820s, following the bankruptcies of private banks in the preceding decade. In 1824 privately owned and operated partnerships with unlimited liability were authorized by royal proclamation and in 1830 the first such note-issuing bank was chartered. Many others followed even though usury laws limited their access to funds. In the 1860s, usury laws were repealed, and joint stock banks were authorized but

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

denied the right of note issue. Private bank notes competed with the Riksbank issues until 1897 when the bank obtained a monopoly of note issues. Private note issues were prohibited and ceased to circulate in 1903. The Swedish banking system was stable, both types of private banks and the Riksbank itself never having been subject to runs or failures. The gold standard and unlimited liability of the private bank issuers promoted, according to Lars Jonung (1984: 363–65), stable note issue policies.

Competitive versus Monopoly Issue in Italy. Proponents of competition in note issue were ascendant when Italy was unified in 1861, at which date nine banks were in existence and various firms and individuals also printed notes. This practice ended in 1874. In 1893 one of the nine banks failed and two others merged, leaving three banks, according to a law of 1894, with the right of issue. Until 1926 this was the extent of competition. It ended with the victory of the Bank of Italy as the sole bank with the right of issue (Fratianni and Spinelli 1984: 408–09).

East Asia and Oceania

Paper Money in China from the 9th to the 20th Centuries. Paper money circulated in China from the 9th century onward, but apparently the issues were made not by private firms but by the rulers of China, who initially maintained reserves of copper cash, and were obligated by law to redeem the emissions triennially. Wars led to overissue, depreciation, and eventually repudiation. By 1567 paper money had ceased to circulate.

During the first years of the reign of the Manchus beginning in 1644, some treasury bills were issued. Otherwise the Manchus issued no paper money until 1852. During the Taiping Rebellion that began that year note issue expanded until it depreciated and became valueless. The government then left note issue to private and provincial banks. These were notably successful in Foochow, the capital of Fukien province, and in other cities, where unregulated local banks issued notes redeemable in copper cash. Notes of larger banks circulated at par, and only 1 large bank of 45 in existence at the conclusion of the episode failed. Failures of smaller banks, however, were more numerous (Selgin 1988: 7–8).

After the Republican revolution of 1911, the new central government favored "modern style" banks, which had given it financial support. These banks issued silver-based notes. When the Nationalists gained power in 1927, they prohibited the issue of copper notes (Selgin 1988: 7–8). Kann (1937: 366–68) reports that overissue

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

occurred in the provinces during the first three decades of the 20th century.

A recent study presents evidence that inside money in China expanded during the years after 1930 while the price level declined and industrial output rose. Despite numerous bank failures, banks by and large maintained convertibility of their liabilities into silver. Although the government had full or partial ownership of three of the largest banks, there was no central bank with a monopoly of bank issues. Other banks possessed the right of issue on the same terms as the largest government-owned bank. Bank notes were backed by 100 percent reserves of silver, gold, foreign exchange, and government securities. There were no reserve requirements for deposits. Adherence to a silver standard in China before November 1935 when a central bank was established is interpreted as a decision of the private banks that operated according to real bills rules. The private banks on this view exemplified free banking under a commodity standard (Brandt and Sargent 1989: 34–38).

In November 1935 the notes of the three largest modern banks were given legal tender status. The Nationalist government monetized its deficits, and the currency became fiat.

Competitive Note Issue in Japan. From the middle of the 16th century onward, feudal lords, merchants, and private cooperatives issued notes. In 1882, when the Bank of Japan became the monopoly issuer, private issues ceased. The record of the success of private issuers in maintaining the value of their notes has not been established (Vaubel 1984: 69).

Competitive Note Issue in Hong Kong. Most Hong Kong notes were issues of four chartered private banks, with government issues limited to one-dollar and subsidiary denominations amounting to less than 10 percent of total issues. The private banks are regulated, and the total authorized issue requires 100 percent reserves of foreign exchange. Excess issues must be covered by certificates of indebtedness of the Hong Kong Government Exchange Fund.

Australia before World War I. Until 1910 several note-issuing banks settled clearing balances with each other in specie. A law then authorized the issue of Australian government legal-tender notes, and in 1911 the Commonwealth Bank was established to act as agent in issuing these notes. A prohibitive 10-percent tax on all private bank note issues was imposed and restrictions on the issue of legal-tender notes were relaxed. In September 1914, when the government declared a gold embargo, the private banks stopped settling balances at the clearinghouse in specie. Thereafter they settled balances in the notes of the monopoly bank of issue (Copland 1920: 490–91).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Conclusion

A summary of the evidence on private competitive banking covering many geographical areas and past monetary systems is difficult. The banks were regulated in a variety of ways. Outside money was usually specie supplemented by government fiat-money elements. A convertibility requirement existed. We cannot be sure that whatever success private competitive banks achieved at different times and places confirms the theory that self-regulation limits overissue.

The Real-Bills Doctrine

Theory

The doctrine maintains that a banking system that confines its lending to discounting short-term self-liquidating commercial bills of exchange arising from real transactions in goods and services—the productive use as opposed to the speculative use of credit—cannot overissue.

Adam Smith was among the first to uphold the real-bills doctrine, although he also insisted on subjecting banks to the legal requirement to convert their notes on demand into specie. Later adherents included the antibullionists who absolved the Bank of England of the charge of overissue during the Napoleonic period when convertibility of its notes into specie was suspended. They argued that, even when inconvertible, overissue of notes was impossible if they were emitted on loans collateralized only by sound, short-term commercial paper. The Banking School, whose chief proponents were Thomas Tooke, John Fullerton, and John Stuart Mill, denied that a convertible currency could be overissued because the needs of trade automatically limited the quantity of money. Moreover, if banks lent on long term or for speculative purposes instead of on real bills only, the law of reflux would cause excess issues to return immediately to banks to repay loans. In the United States the real-bills doctrine was popular with many supporters of the act that created the Federal Reserve System. According to the Act, the amount of Federal Reserve notes to be issued would depend on "notes, drafts, and bills of exchange arising out of actual commercial transactions" (section 13), offered for discount rates to be established "with a view of accommodating commerce and business" (section 14d).

Criticism of the real-bills doctrine has an equally long history. Henry Thornton denied that the quality of the loans securing note issues provided a proper limit on the size of the issues. Ricardo was also a prominent critic. One problem with the doctrine is that even if each loan were made on short-term commercial paper, the volume of bills

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

depends on the turnover of goods in process. The same goods sold a number of times could generate an equal number of bills. The same applies to the term of the bills, which might exceed the period of turnover of the goods. Moreover, it is not clear how banks are supposed to be able to distinguish real from fictitious bills.

Hence the money supply could greatly exceed the needs of trade. The fact that an increase in money supply contributes to a rise in commodity prices would justify a still further increase in money supply to meet the needs of trade. The money value of real transactions therefore cannot serve as an effective regulator of the quantity of money. A further fallacy of the doctrine is its implicit assumption that no one would pay interest on unneeded funds but instead would return such loans to the bank. The doctrine ignores the fact that the loan rate of interest may be below the expected rate of return on funds borrowed. Such a differential encourages borrowing, causing money and prices to rise as long as the interest rate differential exists on discounted real bills (Humphrey 1982: 3–13). In addition, the choice of assets governing the way in which banks introduce their liabilities provides no check on the amount of liabilities banks circulate. In short, the real-bills criterion provides no effective limit on money or prices.

Recently "something of a rehabilitation of the real-bills doctrine" has been claimed (Sargent and Wallace 1982: 1214). The version of the doctrine supposedly rehabilitated, however, bears little resemblance to the actual doctrine. That version is based on a model of an overlapping-generations consumption economy in which agents have perfect foresight, population is exogenous, and there is no production. For this model Sargent and Wallace show that a laissez-faire banking regime leads to price-level instability or indeterminacy. Since real-bills advocates believed that the doctrine guaranteed price-level stability, and its critics argued that it did not, Sargent and Wallace clearly have not rehabilitated the doctrine—even if their version of it were accurate—but have instead confirmed its critics' views. Their version of the doctrine as one characterized by the absence of government restrictions on bank intermediation, however, is a caricature. The actual doctrine was not a defense of laissez faire in banking, but one that mandated specie convertibility of notes on demand and prescribed the class of bills eligible for discount. Moreover, in the Sargent and Wallace model, banks make unlimited consumption loans, whereas according to the actual doctrine banks were to lend only to finance current production and distribution and eschew all other types of lending (Laidler 1984: 149–55).

Another restatement (Smith 1994: 28–29) of the real-bills doctrine "in modern form" asserts that there is a crucial distinction between

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

the consequences of monetary change depending on how the change was accomplished. (It is debatable whether the restatement is "modern," since it was already expressed in 1844 by John Stuart Mill, if not earlier.) On this view what matters is whether the claims purchased in creating money are backed by real goods or a sinking fund. Thus an open-market operation that swaps an asset in private portfolios for government currency is held to have no effect on the price level. Only printing press issues that are not swapped, on this view, have price effects. Monetary changes that raise seigniorage revenue would be expected to be inflationary, and those that are pure portfolio rearrangements would not be. Theoretically, this approach elevates what happens as a result of the first round of spending to encompass the effects of subsequent rounds.

Evidence

We first look at evidence that relates to the traditional version of the real-bills doctrine, then at its "modern form."

Two bits of evidence in the literature give a negative verdict on the traditional real-bills doctrine: Britain during the Napoleonic war-period of suspension of specie payments, and Germany during the 1922–23 hyperinflation. Evidence in favor of the modern form has been offered for the American colonial period, for the Second Bank of the United States, for China, 1930–35, for the ends of four big inflations, and for Taiwan. How valid is the latter evidence?

Britain during Suspension of Specie Payments. From February 1806 to February 1811 the circulation of Bank of England notes rose from £17.7 to 23.4 million, and its accommodation to private borrowers from £11.8 to 19.0 million, while its coin and bullion declined from £6.0 to 3.4 million. The premium on bullion rose from 110.0 in 1806 to 121.1 in 1811. Commodity prices rose 15 percent over the corresponding period. The *Report of the Bullion Committee to Parliament* in 1810 arrayed the bullionists, who criticized monetary developments, against anti-bullionists, who defended the Bank of England and the government. The anti-bullionists claimed that so long as banks issued notes only to discount sound short-term commercial paper overissue was impossible. No one would borrow at interest amounts that he did not need. They attributed inflation to cost-push factors, such as wartime shortages. The bullionists pointed out that the rate of interest banks charged was no regulator of issue when usury laws prevented the Bank of England from charging more than 5 percent. Thornton also showed that with rising prices, the real rate of interest was below the nominal rate and hence made borrowing even more profitable. The anti-bullionists offered no counterargument.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

German Hyperinflation, 1922–23. The Reichsbank pegged its discount rate at 5 percent until the middle of 1922, and then raised it as high as 90 percent by October 1923, when the market rate of interest was a huge multiple of the discount rate. In effect, the bank's rediscounts were cost free to the commercial banks and their debtors. Private Reichsbank credits to industry and commerce increased from .03 billion marks in June 1922 to 1,215 thousand trillion in October 1923, and the sum of Reichsbank and private bank notes was double the credits at the later date. The Reichsbank's president believed that the bank was simply meeting the legitimate needs of business and therefore was not responsible for the hyperinflation (Angell 1929: 368–71).

The evidence that has been adduced in support of the so-called modern form of the real-bills doctrine associates failure of prices to rise following injections of money as symptomatic of asset backing for the new money either through the levy of future taxes or private sector bills. One example of the evidence is the experience of China during the U.S. silver purchase program.

China, 1930–35. The experience of China during this period has been interpreted as a successful demonstration of the working of the real-bills doctrine. In those five years the money supply increased from 6.1 billion to 7.8 billion yuan, but the price index, except for a rise in 1931, fell in every other year. Bank portfolios consisted of government securities and loans to finance inventories. Both are held to be real bills, the government securities because the levy of future taxes provided backing for them, the inventory loans because they were secured by commodities deposited in warehouses of the lending banks or in other warehouses (Brandt and Sargent 1989: 33–47).[2]

The characterization of the Chinese banking regime as one that followed real-bills rules flouts the historical content of the doctrine. No believer in real bills, as the doctrine was formulated, would have accepted the presence of government securities in a bank's portfolio, whatever their backing, as validating the doctrine. It is further questionable that loans on warehoused inventories would have qualified

[2]The context in which Brandt and Sargent introduce real bills is their reinterpretation of the influence on China of the U.S. silver purchase program of 1933. They deny that that program caused internal deflation in China as its currency appreciated, and that the deficit in its balance of payment led to silver exports that contracted the internal money supply. They contend instead that China experienced no economic difficulties and that it was the deliberate choice of the Chinese government to abandon the silver standard in order to acquire the capital gain from silver appreciation and to be free of restrictions on government finance that a commodity standard imposed. For a contrasting view, see Friedman (1992: 157–88).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

as real bills. Inventories more likely would have been regarded as frozen assets, especially when commodity prices were falling.

Conclusion

The real-bills doctrine, in its traditional or restated form, seeks to restrict the definition of overissue. Examples of changes in the quantity of money that are asserted not to have had price effects are cited as validating the theory. The historical examples, however, involve data of uncertain reliability. The emphasis on the source of the initial injection of money is myopic. The sources of the initial injection are not distinguishable in the money stock.

An Alternative Proposal: "Self-Regulation" by Banks, with Owners and Managers at Risk

Theory

The propositions that we have examined in the preceding sections in our judgment are an unsatisfactory basis for the case made by proponents of self-regulation by banks. We offer an alternative approach that we believe makes a better case for deregulation. Banks are not inherently unstable, provided the institutional setting in which banks operate and the incentives they are exposed to do not predispose them to excessively risky undertakings. It is possible to design a banking system in which banks are free to choose any assets yielding the highest return as they judge it and to pay whatever price for funds they deem reasonable, with minimum risk for depositors and creditors. One condition is that the institutional setting fosters price stability. The importance of this condition is related to the procedures banks rely on in making loan and investment decisions. The value of the collateral for secured loans and projections of the ratios of borrowers' current assets to current liabilities for unsecured loans serve as the basis for credit analysis. Unanticipated price-level change can invalidate the assumptions underlying bank portfolios. Price-level stability promotes a stable banking system. We defer to a later section issues related to the achievement of a stable price level.

A sine qua non of a banking system in which the activities institutions engage in are no longer monitored is that a set of restraints be imposed on the conduct of owners and managers. Any mistakes they make that wipe out capital and reserves would prompt the immediate closing or reorganization of the institution. Stockholders might be liable for an assessment equal to their equity, and managers would lose their jobs. Depositors and creditors under these arrangements would be

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

fully protected, as net worth would not have turned negative, should the marginal costs of products and services the institution offers the market turn out to exceed market prices. Whatever risks institutions were prepared to undertake, whether the activities were on or off balance sheets, the mistakes would be borne by the owners and managers, not by depositors and taxpayers. A set of restraints that would safeguard against excessive risk taking might include capital requirements, marking assets to market, provision of information to authorities, and double liability for shareholders.

In this setup, a regulatory agency is needed to certify the integrity of individual institutions. In principle, it could be a private agency, although to have the right to close down an institution would require legislative authorization. Whether quasi-private or governmental, the agency would function under a changed set of conditions, concerned with market-value accounting, capital adequacy, and the information required from the institutions it supervises, not with bank location or branching, or the products and services that banks provided.

The existing regulatory framework encourages regulators to paper over bank failures by relaxing regulatory standards and by giving subsidies to failing institutions. They are responsive to concerns for loss of jobs as a result of bank failures, and the burden imposed on borrowers of establishing credit with a new bank. In the proposed regulatory climate, it would be in the self-interest of regulators not to delay recapitalizing, merging, or liquidating an institution the capital of which had declined to zero. Failed banks that other banks acquire entail no loss of jobs. If shut down, they will be replaced by other banks if there is a market for their services. Timely action by regulators that would safeguard depositors' interests, minimize the risk of disruption of the payments mechanism, and eliminate third-party effects would burnish their image as promoters of the public welfare.

With depositors' funds secure, bank failures would not precipitate runs. The lender of last resort would not need to anticipate panics. If deposit insurance were maintained, it could be privately administered, with the aim of providing depositors and creditors in reorganized or closed institutions immediate access to the full value of their deposits or their claims.

In such a banking system, self-regulation would prevail, in the sense that each institution would have an incentive to be prudent in acquiring assets and funding their acquisition, on penalty that foolhardy owners and managers would lose their equity and stake in the institution. Safe and sound banking would be achieved, contingent on the maintenance by monetary authorities of a stable price level.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Evidence

To identify banking systems that conform to the theory, one must establish in which periods approximate price stability prevailed and whether during those intervals banks essentially were free of restrictions on their choice of activities and provision of inside money. We are not aware that such evidence exists.

Lawrence H. White has argued forcefully that Scottish experience teaches us that government has no role to play in regulating private firms that produce money (1984: 137–50). For the period that he celebrates as validating laissez-faire principles, we lack information on the detailed assets and liabilities of Scottish banks. The record of price movements in Scotland was presumably parallel to that in Britain, since both shared similar experience in adhering to or suspending convertibility. The record was one of price instability.

But even if White could provide detailed evidence supporting superior economic stability in Scotland than in Britain, the question of the applicability of that finding to present-day economic and financial circumstances would remain. He relies on "a very short period [that] typically elapsed between the issue of a bank note and the fulfillment of a promise to pay, the note returning to its issuer through the clearinghouse after deposit in another bank" (1984: 140) to check both fraud and "indeliberate" overissue. Perhaps that was adequate in Scotland before the mid-19th century. We are skeptical that that would be adequate protection currently for note holders and depositors, in an era of global banks engaged in novel activities, given the incomplete information about today's complex world available to economic agents. That is why we can endorse White's plea for free banking in the provision of inside money at the same time that we recognize a need for safeguards. We propose giving the utmost scope for lending and investing by depository intermediaries, without interference by government, while reserving a specialized role for a regulatory agency that may be governmental as judge of institutions' performance. That role is to close or reorganize depository institutions if they have been mismanaged before losses have been sustained by their liability holders.

Whatever risks institutions were prepared to undertake, whether the activities were on or off balance sheets, the mistakes would be borne by owners and managers, not by note holders, depositors, and taxpayers. Bank failures invite government intervention either to bail out management or to make bank depositors and creditors whole, using taxpayer funds. Closing institutions before their net worth has turned negative is a way of precluding government intervention.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Possible Constraints on Outside Money

Economists can offer no well-understood prescription to achieve price-level stability other than limited growth of outside money, that is, the monetary base. Gold and bimetallic standards—commodity-money systems—have the effect of limiting excessive growth of the base. White and Selgin, who favor gold convertibility as the complement to the complete deregulation of inside money and disestablishment of central banks, do not discuss the defects of the gold standard and the absence of professional and political support for it (U.S. Gold Commission 1982).

In our judgment, restoration of either a gold- or bimetallic-based monetary standard is highly unlikely, and we doubt that central banks will cease to function as monetary authorities in the foreseeable future. With few exceptions (Canada, New Zealand), each of them now essentially operates a fiat monetary standard at its discretion, with no commitment to maintain any particular price level or rate of inflation. Under existing monetary arrangements, to the extent that banks make contracts that stipulate fixed rates of future money payments to others and from others, the consequences are arbitrary. This is so because banks have no way of knowing whether the price level or the inflation rate they expect will be realized. What is needed is a formula for central banks to follow in creating outside money that will limit its growth, with the objective of approximating a long-run stable price level.

A quarter of a century ago Milton Friedman advocated a constant rate of money growth at 4 percent a year (assuming real income growth of 3 percent a year and a 1 percent annual decline in velocity) to achieve a zero inflation rate. Although convinced that the consequences of this rule would have been preferable to the discretion that created first inflation and then disinflation, for two reasons Friedman has since abandoned this strategy. The variable the monetary authorities control directly is outside money. Their control of inside money is indirect and from quarter-to-quarter is loose. The particular numbers Friedman relied on also have not been constant over time. He believes that his suggestion has encouraged the Federal Reserve to plead inability to live by the rule and that it has thus escaped accountability. He then suggested a constitutional amendment that would establish a range of 3 to 5 percent a year in the growth of outside money, but discarded it as a half-measure of reform. As a full-measure, he proposed instead that, after a transition period, outside money be frozen at a fixed amount. Private depository institutions would then be freed to issue bank notes. The money-creating powers of the central bank would cease (Friedman 1984: 40–52).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

In choosing the money stock for his initial rule, Friedman's objective was to link that magnitude to subsequent changes in national income. In the spirit of Friedman's suggestion, Bennett McCallum (1989: 329–43) has proposed a rule linking the quarterly change in outside money to the quarterly change in nominal GNP. His rule, he concludes, would lead to roughly 3 percent annual growth in nominal GNP and approximately zero inflation.

Whether monetary authorities will set aside conflicting objectives that have made their policies erratic and unpredictable and be resolute in pursuing price stability as their overriding concern, no one can say. If they fail to achieve price stability, the performance and stability of banking systems under "self-regulation" will be in jeopardy.

Conclusion

We have surveyed three theories and the evidence that is said to confirm that under private market arrangements self-regulated banking systems will limit their inside money issues. These theories argue that intermediation by note- and deposit-issuing banks can safely be deregulated because:

1. An interbank clearing mechanism checks overissue;
2. Competitive producers of money establish confidence in their issues by limiting their magnitude;
3. Banks that restrict their discounting to real bills cannot overissue.

Each of the theories is problematic. The supporting evidence is drawn from many countries and periods in the past. In no episode is it possible to isolate the particular theoretical proposition it is supposed to confirm. Bank performance may have been influenced by existing regulations and the requirement that inside money be convertible into some form of outside money. The significance of the theoretical proposition is accordingly ambiguous. Moreover, the relevance of historical banking experience to current economic and financial conditions is overdrawn.

Our support for deregulation of bank intermediation features two important provisos: First, the risks that are undertaken are borne by owners and managers, not by money holders and taxpayers, and second, monetary authorities can be relied on to achieve price-level stability.

The first proviso leads us to advocate restraints on depository institutions such as capital requirements, provision of information to a supervisory agency, marking assets to market, and possibly double liability for equity holders. The key responsibility of the regulatory agency

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

would be to close or reorganize promptly an institution before its net worth turned negative, thus safeguarding the funds of depositors and creditors. The agency could be administered as a private undertaking if the legislature empowered it to close troubled institutions.

The second proviso leads us to inquire about possible ways to constrain the issue of outside money. A return to a commodity-based standard by the world seems improbable. We review some suggestions to limit the issue of fiat outside money, the only prescription economists can offer to achieve price-level stability. Price-level stability in our view is a prerequisite for safe and sound banking. If the price level is stable, any mistakes that banks make in acquiring assets will be attributable to faulty credit analysis, not to price-level or inflation surprises. If the outlook for future price-level stability is doubtful, the success of a self-regulated banking system is also doubtful.

References

Angell, J.W. (1929) *The Recovery of Germany*. New Haven, Conn.: Yale University Press for the Council on Foreign Relations.

Bordo, M.D., and Redish, A. (1987) "Why Did the Bank of Canada Emerge in 1935?" *Journal of Economic History* 47(2): 405–17.

Brandt, L., and Sargent, T.J. (1989) "Interpreting New Evidence about China and U.S. Silver Purchases." *Journal of Monetary Economics* 23(1): 31–51.

Carr, J., and Mathewson, F. (1988) "Unlimited Liability as a Barrier to Entry." *Journal of Political Economy* 96(4): 766–84.

Checkland, S.G. (1975) *Scottish Banking: A History, 1695–1973*. Glasgow: Collins.

Cooper, R.N. (1989) "Is Private Money Optimal?" *Cato Journal* 9(2): 393–97.

Copland, D.B. (1920) "Currency Inflation and Price Movements in Australia." *Economic Journal* 30(4): 484–509.

Dewey, D.R. (1910) *State Banking before the Civil War*. National Monetary Commission. 61st Cong., 2nd Sess. Senate Doc. No. 581. Washington, D.C.: U.S. Government Printing Office.

Dowd, K. (1989) *The State and the Monetary System*. New York: St. Martin's Press.

Fraas, A. (1974) "The Second Bank of the United States: An Instrument for an Inter-regional Monetary Union." *Journal of Economic History* 34(2): 447–67.

Fratianni, M., and Spinelli, F. (1984) "Italy in the Gold Standard Period, 1861–1914." In Bordo, M.D., and Schwartz, A.J. (eds.) *A Retrospective on the Classical Gold Standard: 1821–1931*, 405–51. Chicago: University of Chicago Press.

Friedman, M. (1984) "Monetary Policy for the 1980s." In Moore, J.H. (ed.) *To Promote Prosperity: U.S. Domestic Policy in the Mid-1980s*, 40–52. Stanford, Calif.: Hoover Institution.

Friedman, M. (1992) *Money Mischief: Episodes in Monetary History*. New York: Harcourt Brace Jovanovich.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Goodhart, C.A.E. (1987) "Review of 'Free Banking in Britain' by L.H. White." *Economica* 54(1): 129–31.
Gorton, G., and Mullineaux, D.J. (1987) "The Joint Production of Confidence: Endogenous Regulation and 19th-Century Commercial-Bank Clearinghouses." *Journal of Money, Credit and Banking* 19(4): 457–68.
Holdsworth, J.T. (1910) *The First Bank of the United States*. National Monetary Commission. 61st Cong. 2nd Sess. Senate Doc. No. 571. Washington, D.C.: U.S. Government Printing Office.
Humphrey, T.M. (1982) "The Real-Bills Doctrine." Federal Reserve Bank of Richmond *Economic Quarterly* 68(5): 3–13.
Jonung, L. (1984) "Swedish Experience under the Classical Gold Standard, 1873–1914." In Bordo, M.D., and Schwartz, A.J. (eds.) *A Retrospective on the Classical Gold Standard, 1821–1931*, 361–404.
Kahn, J.A. (1985) "Another Look at Free Banking in the United States: New Evidence on the Free-Banking Era." *American Economic Review* 75(4): 881–85.
Kann, E. (1937) "The Currencies of China: Old and New." In Gayer, A.D. (ed.) *The Lessons of Monetary Experience: Essays in Honor of Irving Fisher*, 363–78. New York: Farrar & Rinehart.
Klein, B. (1974) "The Competitive Supply of Money." *Journal of Money, Credit and Banking* 61(4): 423–53.
Laidler, D. (1984) "Misconceptions about the Real-Bills Doctrine: A Comment on Sargent and Wallace." *Journal of Political Economy* 92(1): 149–55.
Landmann, J. (1910) *The Swiss Banking Law*. National Monetary Commission. 61st Cong. 2nd Sess. Senate Doc. No. 401. Washington, D.C.: U.S. Government Printing Office.
McCallum, B.T. (1989) *Monetary Economics: Theory and Policy*. New York: Macmillan.
Mullineaux, D.J. (1987) "Competitive Money and the Suffolk Bank System: A Contractual Perspective." *Southern Economic Journal* 53(4): 884–98.
Munn, C.W. (1985) "Review of 'Free Banking in Britain' by L.H. White." *Business History* 27: 341–43.
Munn, C.W. (1991) "Comment on L.H. White: 'Banking without a Central Bank'." In Capie, F., and Wood, G.E. (eds.) *Unregulated Banking: Chaos or Order?*, 63–67. London: Macmillan.
Ng, K. (1988) "Free-Banking Laws and Barriers to Entry in Banking, 1838–1860." *Journal of Economic History* 48(4): 877–90.
Rockoff, H. (1975) *The Free-Banking Era: A Reconsideration*. New York: Arno.
Rockoff, H. (1991) "Lessons from the American Experience with Free Banking." In Capie, F., and Wood, G.E. (eds.) *Unregulated Banking: Chaos or Order?* 73–109.
Rolnick, A.J., and Weber, W.E. (1983) "New Evidence on the Free-Banking Era." *American Economic Review* 73(5): 1080–91.
Rolnick, A.J., and Weber, W.E. (1984) "The Causes of Free Bank Failures: A Detailed Examination." *Journal of Monetary Economics* 14(3): 267–91.
Rothbard, M.N. (1988) "The Myth of Free Banking in Scotland." *Review of Austrian Economics* 2: 230.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Sargent, T.J., and Wallace, N. (1982) "The Real-Bills Doctrine versus the Quantity Theory: A Reconsideration." *Journal of Political Economy* 90(6): 1212–36.

Schuler, K. (1992) "The World History of Free Banking." In Dowd, K. (ed.) *The Experience of Free Banking*, 7–47. London: Routledge.

Schweikart, L. (1991) "U.S. Commercial Banking: A Historiographical Survey." *Business History Review* 65 (Autumn): 606–61.

Sechrest, L.J. (1988) "White's Free-Banking Thesis: A Case of Mistaken Identity." *Review of Austrian Economics* 2: 251–52.

Selgin, G.A. (1988) *The Theory of Free Banking: Money Supply under Competitive Note Issue*. Cato Institute. Totowa, N.J.: Rowman & Littlefield.

Smith, B.D. (1994) "Mischief and Monetary History: Friedman and Schwartz Thirty Years Later." *Journal of Monetary Economics* 34(1): 27–45.

U.S. Gold Commission (1982) Report to the Congress of the Commission on the Role of Gold in the Domestic and International Monetary Systems. Washington, D.C., 31 March.

Vaubel, R. (1984) "A History of Currency Competition." In Salin, P. (ed.) *Currency Competition and Monetary Union*. The Hague: Martinus Nijhoff.

Weber, E.J. (1988) "Currency Competition in Switzerland, 1826–1850." *Kyklos* 41(3): 459–78.

White, E.N. (1991) "Experiments with Free Banking during the French Revolution." In Capie, F., and Wood, G.E. (eds.) *Unregulated Banking: Chaos or Order?*, 131–50.

White, L.H. (1984) *Free Banking in Britain: Theory, Experience, and Debate, 1800–1845*. Cambridge: Cambridge University Press.

White, L.H. (1989) "What Kinds of Monetary Institutions Would a Free Market Deliver?" *Cato Journal* 9(2): 367–91.

White, L.H. (1991) "Banking without a Central Bank: Scotland before 1844 as a 'Free-Banking' System." In F. Capie, and Wood, G.E. (eds.) *Unregulated Banking: Chaos or Order?* 37–62.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Bank Self-Regulation: Comment on Bordo and Schwartz

George Selgin

Can banking systems regulate themselves? Hardly anyone thought so a decade ago. Since then numerous writings (see Selgin and White 1994) have defended the opposite view. Critical responses to these writings serve an important purpose, by spelling out the precise nature of supposed market failures in money and banking—failures that are often simply taken for granted.

The Bordo-Schwartz critique of bank self-regulation focuses on three propositions offered in support of monetary laissez faire. These are:

1. That interbank clearings are capable of preventing the overissuance of bank money (notes or deposits) both by individual banks and by the banking system as a whole;
2. That competing, private issuers of paper currency will deliberately aim to stabilize the purchasing power of their monies; and
3. That the quantity of money will automatically be limited if banks restrict themselves to discounting "real bills."

Bordo and Schwartz argue that all three propositions are false, and that the possibility of bank self-regulation is therefore "doubtful."

Visions of Monetary Laissez Faire

It is important to recognize that each of the three propositions listed above is linked to a distinct theory or "vision" of monetary self-regulation. The propositions are thus independent from if not at odds with one another. Proposition (1), concerning the clearing mechanism, is appealed to by proponents, including White (1984) and myself (1988), of "free banking," where commercial banks freed from government regulations are still assumed to redeem their notes and deposits

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

The author is Associate Professor of Economics at the University of Georgia. He wishes to thank his colleagues Chris Cornwell and Dick Timberlake for their help and suggestions.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

in some base money produced outside the banking system. Proposition (2), in contrast, is mainly appealed to by proponents of competing, private fiat-type monies, including Klein (1974), and Hayek (1978). Proposition (3), finally, is potentially applicable to any sort of banking system, whether free or centralized. It is associated with only one important, recent theoretical paper, written by Thomas Sargent and Neil Wallace (1982).

It follows that no single proponent of monetary self-regulation should be expected to subscribe to, much less defend, all three propositions. However, as I point out later, Bordo and Schwartz take for granted that any argument for monetary self-regulation, including arguments for free banking, must rest on all three of the above propositions. They therefore mistakenly assume that theory and empirical evidence that contradicts the second and third propositions also undermines the case for free banking. That simply is not so: the modern case for free banking rests in part (though not entirely) on proposition (1) above, and not at all on propositions (2) or (3). In fact, in my own writings (Selgin 1989; Selgin and White 1994: 1734–36) I explicitly deny the validity of the latter propositions.

Self-Regulation through the Clearing Mechanism

The free-banking vision of monetary laissez faire is one in which competing banks, free from all government regulations other than ones governing the mass of business contracts, offer deposits (D) and banknotes (N) that are redeemable in some base money produced outside the private banking system. The quantity of money supplied by such a system depends on (1) the stock of base money (B), all of which is assumed to be kept in the form of bank reserves (R), and on (2) the banking-system reserve ratio (r), where $D + N = B(1/r)$. This value will, of course, be well defined only in so far as both B and $1/r$ (the free-banking money multiplier) have well-defined values.

All modern discussions of free banking acknowledge a need for institutional arrangements capable of restricting the stock of base money. Some take for granted a commodity (e.g., gold) standard, where base money is subject to increasing average real production costs. Others (e.g., Selgin 1988: chap. 11; Selgin 1994b) allow for a fiat standard subject to a strict quantity rule, such as a frozen monetary base. Given some such arrangement for limiting base growth, asking whether a free banking system can effectively restrain the growth of any monetary aggregate is equivalent to asking whether the free-banking money multiplier is determinate. Affirmative answers to this

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

question point to the bank money clearing mechanism as a multiplier-constraining device.

In response Bordo and Schwartz argue that, although the clearing mechanism is capable of preventing individual banks from issuing too many notes or deposits relative to other banks in the system, the same mechanism cannot prevent a *concerted* overexpansion: if all banks in a system expand in unison, none suffers any net reserve loss. The money multiplier will therefore be indeterminate.

This is a venerable argument against unregulated banking, whose past proponents include both John Maynard Keynes and Chicago economist Lloyd Mints. Nevertheless the argument is theoretically suspect, because it ignores the role bank reserves play in protecting banks against random reserve losses. As long as interbank transactions involve a random element, then (assuming positive costs of liquidating non-reserve assets on short notice) any individual bank will want to hold some reserves to guard against the positive probability of suffering an adverse clearing in any single clearing session. Even if net clearings are expected to zero-out eventually for all banks, a bank starting out with zero reserves would face 50:50 odds of being in default after a single clearing session. To reduce these odds to something more reasonable, the bank will want to hold precautionary reserves. The more reserves it starts with, the lower its risk of running short.

More crucially, the greater the volume of clearings, as given by the total stock of bank money times its velocity of circulation, the greater will be the reserve ratio needed by an individual bank to achieve any given level of protection from default.[1] Assuming, then, that a group of banks starts out with some "optimal" level of protection and that the stock of bank reserves is given, any in-concert expansion will prove unsustainable, because it will involve a reduction in bank reserve ratios below their optimal level.

Bordo and Schwartz acknowledge this argument,[2] but dismiss it as a "rudimentary" concoction of one or two proponents of free banking. In fact the argument is simply an application of the standard (and, as far as I am aware, still unsurpassed) microeconomic theory of bank precautionary reserve demand, which is itself merely a version of the received theory of precautionary money holding. Bordo and Schwartz

[1]The argument therefore does not assume that the free-banking reserve ratio is fixed, as Bordo and Schwartz claim.

[2]At one point, however, they assert that free-banking theorists have not "understood that a clearinghouse provides no check if all banks expand simultaneously," wrongly implying that we have not been aware of the in-concert overexpansion argument.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

are therefore unjustified in rejecting the argument out of hand, without presenting any counterargument save for the claim that I personally have not employed the argument consistently.[3]

Historical Evidence

Instead of attempting to answer the theoretical argument for the clearing mechanism as a mechanism for bank self-regulation, Bordo and Schwartz review some historical evidence bearing on the subject. The main episodes they consider are (1) Scottish banking before 1844; (2) the New England Suffolk System; (3) the Second Bank of the United States; and (4) Canadian banking before 1914. They claim that the evidence on the whole contradicts claims concerning the effectiveness of the clearing mechanism in restraining aggregate bank expansion.

Just how damaging is this "negative" evidence to the case for free banking? The most damaging evidence would appear to be that for the Second Bank of the United States. It is generally agreed that the Second Bank led a systemwide overexpansion of currency from 1828 to 1831. What is controversial is Bordo and Schwartz's claim that this episode illustrates the consequences of free banking. Indeed, many writers regard the Second Bank of the United States as an early central bank. While this may be stretching the point, the fact remains that the Second Bank enjoyed many special privileges, including the ability to branch nationwide, and that these privileges together with political incentives faced by the Second Bank caused a breakdown of the bank-money clearing mechanism. Instead of redeeming notes issued by the Second Bank, state banks held and reissued them. The Second Bank, in turn, often refrained from redeeming notes issued by state banks, believing that such a policy would help it to secure political support needed for a renewal of its charter. Such a "live and let live" arrangement was a far cry from what took place in other, less hierarchical banking systems, where special privileges and politics did not undermine banks' normal incentives to compete aggressively with one

[3]The reference is to my discussion (1988: 150–51) of overexpansion in the post-WWI Australian monetary system. In fact, I am not inconsistent. The Australian system in question operated within the confines of an international gold standard. In it, a general overexpansion of bank money (relative to the rest of the world as well as to the domestic demand for real money balances) could occur as a result of capital inflows or domestic gold production, both of which would inflate Australian bank reserves. This sort of overexpansion does not involve any weakening of bank reserve ratios or extension of the bank-money multiplier, and so would not be constrained by the clearing mechanism. Instead, international equilibrium would evolve by way of the price-specie-flow mechanism, involving a net transfer of gold from Australia to other gold-standard nations.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

another. It is to such arrangements, and not to the United States prior to 1837, that one must look for evidence for or against bank self-regulation.

One such arrangement, the New England Suffolk system, was hierarchical in that its clearinghouse was organized and operated by the Suffolk Bank of Boston. But it was not hierarchical to the point of fostering a "live and let live" arrangement between the Suffolk and other New England banks. On the contrary: the system promoted the active clearing and redemption of all bank notes, with close to $400 million in redemptions overseen by the Suffolk bank in 1857 alone.

Bordo and Schwartz believe, however, that evidence from the Suffolk system "contradicts the theory of 'self-correcting in-concert expansions'." The evidence they refer to consists of figures for Suffolk system redemptions, which, Bordo and Schwartz observe, exhibit both cyclical behavior and a "clear upward trend."

Variations in total redemptions alone, however, neither contradict nor support the claims of free-banking theory. That theory implies, not that the volume of redemptions will be stable in any absolute sense, but that it will be a stable function of several other variables, the most important of which are the stock of bank reserves, real income, and the number of banks participating in the redemption arrangement (Selgin 1994a). A cursory glance at statistics representing these other variables suggests that much of the variation in Suffolk redemptions can be attributed to their movements. For example, both real (textile) output and specie holdings of New England banks doubled during the period in question. The number of New England banks also doubled.[4] Allowing for reserve economies of scale, these facts may account for the near quadrupling of Suffolk redemptions. The available statistics are, however, crude proxies only: more careful research will have to be undertaken before any reliable conclusions can be drawn concerning the effectiveness of the Suffolk system in constraining the aggregate stock of New England bank money. For the time being, the evidence remains inconclusive.

"Inconclusive" is also the verdict Bordo and Schwartz themselves reach with regard to evidence from Scotland and Canada. What they mean is that, although both arrangements appear to have been conducive toward overall monetary stability, it is not clear that the clearing mechanism was responsible for this stability. Perhaps, as in the case of New England, no sufficiently detailed statistical analysis of Scottish bank clearings exists to either confirm or refute the clearing-mecha-

[4]Banking statistics are from Root (1895); New England textile output figures are from Davis and Stettler (1966: 221).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

nism hypothesis. Nevertheless, that hypothesis still provides *part* of the best available explanation for the stability of the Scottish and Canadian arrangements. Other explanations, including those offered by Bordo and Schwartz, are far less adequate: they fail either because they emphasize mechanisms common to many less successful banking systems (e.g., the constraint provided by the price-specie-flow mechanism) or because they emphasize mechanisms that were not truly characteristic of successful ones (e.g., unlimited liability, which was not a feature of the three largest Scottish banks).

I emphasize "part" in the last paragraph because it is by no means the case that the clearing mechanism is the only important source of stability in a free-banking system. Other features of past free-banking systems, such as the gold standard, unrestricted branch banking, competitive note issuance, and convertibility were also important, both because they contribute to the efficiency of the clearing mechanism and for other reasons. Although it is perfectly appropriate to observe, as Bordo and Schwartz occasionally do, the positive role played by these other features of past, successful banking systems, it is wrong to imply that this contribution is somehow at odds with the claims of free-banking theory.

Competing Fiat-Type Monies

Free banking is an arrangement in which unregulated private banks issue notes and deposits that are generally redeemable on demand for some "outside" base money.[5] A tendency exists for all banks to redeem their notes in a common base money, such as gold coin; and in the absence of restrictions on branching like-denomination notes issued by different banks all tend to circulate at par:[6] in other words,

[5]In some instances, notes and deposits may also bear an "option clause" granting issuers the right to suspend payment under certain prespecified conditions.

It should be noted that the term "free banking" is also used to refer to banking systems organized in the antebellum United States under so-called free-banking laws. These systems were, however, far from being truly free from significant regulations: most prohibited branch banking, and all required that bank notes be backed by specific assets, typically state government bonds.

[6]See Selgin and White (1987). Bordo and Schwartz appear to accept Cooper's (1989: 393–94) argument that government laws making particular bank notes publicly receivable would tend to undermine par-circulation of multiple private bank notes. I know of no case where this has occurred, and believe that both Canada and Sweden offer counterexamples. Of course, other kinds of government interference can and historically have undermined par-circulation of bank notes. U.S. laws against branch banking are a conspicuous example. But why construe this as an argument against free banking rather than as an argument against the laws in question?

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

various "brands" of bank currency trade at fixed rather than floating exchange rates.

A literature distinct from the literature on free banking imagines a very different arrangement in which competing private firms issue distinct, *irredeemable* fiat-type monies that trade at floating exchange rates, much like today's government-issued fiat monies. According to Hayek (1978) and Klein (1974), competitive pressures would encourage firms to stabilize their monies' purchasing power. Because private fiat-type monies are not claims to any "outside" reserve medium, their issuers do not have to hold any liquid reserve assets, and there is no need for a bank-money clearing mechanism. Overissuing institutions are disciplined, not by adverse clearings, but by a loss of market share to other, more conservative firms.

After briefly reviewing the theory of private fiat-type monies, Bordo and Schwartz rightly question whether competition among private fiat-type monies could be or ever has been observed in practice. They then generate confusion, however, by turning once again to review various historical episodes which, for the most part, have little bearing upon the theory being criticized: although the episodes Bordo and Schwartz refer to in this section of their paper all involve multiple note-issuing institutions, among them only the New England colonial episodes and Italy in the last half of the 19th century experimented with competing fiat monies. These experiments were, moreover, a far cry from laissez faire (Sannucci 1989). Most of the other episodes referred to (Scotland before 1845; Switzerland throughout most of the 19th century; Sweden before 1897; France from the revolution to 1806; mainland China before the 20th century; and Australia before 1910) are (approximate) instances of free banking, not competing fiat monies. Still other episodes, including the misnamed U.S. "free-banking" era, Japanese banking before 1882, and Hong Kong earlier this century, involved neither free banking nor competing fiat monies, and so are not particularly useful for assessing either theory. Indeed, it is rather misleading for Bordo and Schwartz to include them in what purports to be a survey of "commonly cited" examples of competing monies.[7]

Evidently most of the episodes reviewed do not provide any support for the competing fiat-monies theory of monetary self-regulation. But Bordo and Schwartz regard this fact as an argument against, not just the Klein-Hayek theory, but all theories of monetary self-regulation,

[7]Although Schuler (1992) includes all of these episodes in his survey of free-banking episodes, that survey covers all major instances of multiple note issuance, including episodes involving major regulatory restrictions.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

including free-banking theory. Thus, in considering again the Scottish episode, Bordo and Schwartz observe, correctly, that it "clearly does not conform to the theory of flexible exchange rates among distinctive monies." They then mistakenly characterize this feature of Scottish banking as one of several "deviations from a free-market system." The truth is that a lack of floating exchange rates only represents a deviation from the competing fiat-monies theory of monetary self-regulation. It does not at all represent a deviation from free-banking theory, which actually predicts fixed exchange rates between various bank currencies within a single base-currency area.

What about Scotland's other alleged deviations from laissez faire? Lawrence H. White (1991) addressed them at length, concluding in each case either that the allegations themselves are not true or that the alleged deviations do not contradict free-banking theory in any crucial way. Regrettably, Bordo and Schwartz seem to have overlooked this. One allegation White does not address is Checkland's claim (1975: 186) that Scottish banks did not consistently offer to cash notes for the public on demand. Bordo and Schwartz argue that this practice, although it may (as Checkland himself claims) have lent stability to the Scottish system, was inconsistent with free-banking theory. But this is not so: the redemption of bank notes over the counter plays no crucial role in free-banking theory; my own analysis actually abstracts from such redemptions altogether, relying solely on interbank settlements as a source of restraint. Indeed, Checkland's observation, if valid, only serves to undermine Bordo and Schwartz's earlier claim that public redemptions of banknotes alone (and not the clearing mechanism) are able to restrain systemwide overexpansion.

At the end of their survey, Bordo and Schwartz again observe that the evidence does not "confirm the theory that self-regulation limits overissue." But the question is not whether the theory is "confirmed" or not but whether it is the best theory available to account for the fact that some of history's least-regulated banking systems have been exceptionally stable and crisis-free:[8] Of course much more research, and empirical research especially, is needed. Nevertheless, I believe that free-banking theory provides what is, up to now, the best available means for explaining a historical record that blatantly contradicts conventional beliefs concerning the inherent instability of unregulated banking.

[8]In Selgin (1994a) I argue, using evidence gathered largely by Bordo and Schwartz themselves, that the freest banking systems have also been among those suffering the fewest banking crises.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The Bordo-Schwartz Alternative

After declaring arguments and evidence for the stability of free banking insufficient, Bordo and Schwartz suggest an alternative scheme for bank self-regulation. Underlying the scheme are three claims concerning requirements for stability:

1. Price stability (meaning a stable index of consumer prices) is essential to stable banking;
2. Any bank that exhausts its capital and reserves should be immediately closed or reorganized; and
3. All risk of loss should be borne by bank owners and managers rather than by bank depositors, note holders, and other creditors.

To satisfy the first requirement, Bordo and Schwartz favor a monetary-base rule. To satisfy the second, they recommend market-value accounting, and the elimination of subsidies to failed banks. To satisfy the third, they recommend that banks be subjected to statutory capital requirements and double liability, among other things.

In evaluating Bordo and Schwartz's scheme as an alternative to free banking, it is first necessary to consider the extent to which it actually differs from free banking. Consider first the goal of price stability. Most historical free-banking systems were based upon a gold standard; and, as Bordo himself (1984) has observed, while the gold standard appears to have allowed considerable short-run price movements, it nonetheless produced much greater long-run price stability than any modern fiat standard.

Bordo and Schwartz believe, however, that, regardless of its price-stabilizing capabilities, a gold standard "is highly unlikely." They therefore propose that price stability should be achieved by way of some "formula for central banks to follow in creating outside [base] money." Their favorite formula appears to be Milton Friedman's frozen monetary-base rule, which places a complete moratorium on central-bank lending and open-market operations. Bordo and Schwartz see this rule as an alternative to free banking because, in their view, proponents of free banking including myself "favor gold convertibility as the complement to complete deregulation of inside [i.e., bank-issued] money."

In fact, I recommended a free-banking reform based on a frozen stock of fiat money several years ago (Selgin 1988: chap. 11). More recently I discussed the implications of free-banking arrangements based on other rule-bound fiat regimes (1994b). The Bordo-Schwartz monetary-base rule "alternative" to free banking is therefore no alternative at all but rather a component of at least one proposed free-

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

banking reform. That much of the free-banking literature assumes a commodity standard—rather than a fiat-base money—reflects, not necessarily a normative preference, but this literature's attempt to shed light on past free-banking episodes that were in fact based on commodity standards.

Next consider Bordo and Schwartz's second requirement, that banks exhausting their reserves and capital should be closed immediately. This requirement also seems perfectly consistent with free banking: after all, government regulations rather than market forces have been the source of that "forbearance" that permits bankrupt financial institutions to go on functioning as if they were still solvent.[9]

Bordo and Schwartz's third requirement—that only bank owners and managers and not deposit and noteholders or taxpayers should bear any risk of loss in connection with bank failures—does represent a genuine departure from free banking. To be sure, bank owners and managers of banks would occasionally suffer losses under free banking, while taxpayers as such would not. But deposit and noteholders also might suffer losses, as was the case in several historical instances of free banking. Free banking therefore does not satisfy Bordo and Schwartz's third, alleged requirement for stability.

Which, I believe, is a good thing, because the requirement in question is fundamentally misguided. Its satisfaction would entail removal of any tendency for consumers to monitor and discipline ill-managed banks: a consumer who bears no risk of loss from bank failures is a consumer who has no reason to care whether his or her bank is likely to fail or not. Such a consumer will be inclined to patronize banks that make risky investments promising high returns. Bankers in turn would then have no choice but to heed the wishes of risk-immune consumers or lose their business entirely. This, I should think, was the hard lesson taught by recent experience with government deposit insurance: a banking system whose creditors bear no risk is likely to become a risky banking system.

Indeed, the principle that "creditors should suffer no losses" is so misguided that no one has ever even thought of applying it to any

[9]Bordo and Schwartz admit that market forces alone may have posed an adequate restraint upon bank mismanagement in Scotland before 1845. However, they "are skeptical that [similar market forces alone would offer] adequate protection currently for noteholders and depositors, in an era of global banks engaged in novel activities, given the incomplete information about today's complex world available to economic agents." Granting that the world today is more "complex" than it used to be (whatever that means), and that information concerning it is "incomplete," I still fail to see how this adds up to an argument for regulation. If it does, then it is no less an argument for regulating all sorts of international economic activities.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

business other than banking. Why should banking be different? The reason typically given is that unregulated banks, unlike other firms, are subject to runs triggered by depositors' and noteholders' desire to avoid failure-related losses: redemption requests are serviced on a "first-come-first-served" basis, so that panic-stricken liability holders have a better chance of recovering their savings if they get to their banks right away. Runs on isolated banks might not be a problem. However, it is widely believed—and Bordo and Schwartz appear to accept the belief—that bank runs become "contagious," spreading from bank to bank, attacking (pre-run) solvent firms as well as insolvent ones, and threatening a collapse of the entire payments system. Such contagious bank runs constitute an "external effect," implying market failure and the need for extraordinary regulatory measures.

Empirical evidence supporting the contagious bank-run hypothesis is, however, quite limited (Kaufman 1994). The evidence is, moreover, particularly scanty for relatively free banking systems (Selgin 1994a). The view that bank runs are contagious appears, in other words, to be more an article of faith among proponents of bank regulation than an impression confirmed by experience.

One final, critical comment on the Bordo-Schwartz proposal seems in order. Unlike free banking, their proposal involves a government-sponsored regulatory agency.[10] The risk exists that any such agency will be subject to political pressures, including pressure from special interests within the banking industry that may ultimately capture the agency. Thus, while it may be naive for proponents of free banking to pretend that complete deregulation and the adoption of a gold (or, for that matter, frozen-fiat) standard is possible, it seems at least equally naive for proponents of regulation to assume that politically based regulatory agencies can be relied upon to cater to consumers' long-run interests.

References

Bordo, M.D. (1984) "The Gold Standard: Myths and Realities." In Siegel, B.N. (ed.) *Money in Crisis*, 197–237. San Francisco: The Pacific Institute.

Checkland, S.G. (1975) *Scottish Banking: A History, 1695–1973.* Glasgow: Collins.

[10]Bordo and Schwartz insist that the agency "could be administered as a private undertaking." But by this they presumably mean merely that it could be self-funding and privately owned. Presumably they do not mean that it would function on a strictly voluntary basis: all banks would have to conform to its regulations, and could be forced to close at its command. To describe such an agency as "nongovernmental" is, of course, highly misleading: according to the above definition, many central banks, including the Federal Reserve System, are "private." They are also indisputably "governmental."

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Cooper, R.N. (1989) "Is Private Money Optimal?" *Cato Journal* 9(2): 393–97.
Davis, L.E., and Stettler, H.L. III (1966) "The New England Textile Industry, 1825–60: Trends and Fluctuations." In *Output, Employment, and Productivity in the United States after 1800*, 213–38. New York: National Bureau for Economic Research.
Kaufman, G. (1994) "Bank Contagion: A Review of the Theory and Evidence." *Journal of Financial Services Research* 8(2): 123–50.
Hayek, F.A. (1978) *Denationalisation of Money: The Argument Refined*. London: Institute of Economic Affairs.
Klein, B. (1974) "The Competitive Supply of Money." *Journal of Money, Credit, and Banking* 6(4): 423–53.
Root, L.C. (1895) "New England Bank Currency: Current Redemption Developments." *Sound Currency* 11(13): 253–84.
Sannucci, V. (1989) "The Establishment of a Central Bank: Italy in the 19th Century." In de Cecco, M., and Giovannini, A. (eds.) *A European Central Bank? Perspectives on Monetary Unification after Ten Years of the EMS*, 244–80. New York: Cambridge University Press.
Sargent, T.J., and Wallace, N. (1982) "The Real-Bills Doctrine versus the Quantity Theory: A Reconsideration." *Journal of Political Economy* 90(6): 1212–36.
Schuler, K. (1992) "The World History of Free Banking: An Overview." In Dowd, K. (ed.) *The Experience of Free Banking*, 7–47. London: Routledge.
Selgin, G. (1988) *The Theory of Free Banking: Money Supply under Competitive Note Issue*. Totowa, N.J.: Rowman and Littlefield.
Selgin. G. (1989) "The Analytical Framework of the Real-Bills Doctrine." *Journal of Institutional and Theoretical Economics* 145(3): 489–507.
Selgin, G. (1994a) "Are Banking Crises Free-Market Phenomena?" *Critical Review* 8(4): 591–608.
Selgin, G. (1994b) "Free Banking and Monetary Control." *Economic Journal* 104(4): 1449–59.
Selgin, G., and White, L.H. (1987) "The Evolution of a Free-Banking System." *Economic Inquiry* 25 (July): 439–57.
Selgin, G., and White, L.H. (1994) "How Would the Invisible Hand Handle Money?" *Journal of Economic Literature* 32(4): 1718–49.
White, L.H. (1984) *Free Banking in Britain: Theory, Experience, and Debate, 1800–1845*. Cambridge: Cambridge University Press.
White, L.H. (1991) "Banking without a Central Bank: Scotland before 1844 as a 'Free-Banking' System." In Capie, F., and Wood, G.E. (eds.) *Unregulated Banking: Chaos or Order?* 37–62. London: Macmillan.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Community Banking, Monitoring, and the Clinton Plan

Ralph Chami and Jeffrey H. Fischer

One of President Clinton's campaign promises was to increase the availability of credit in poor urban areas by establishing 100 community development banks throughout the United States. The banks would "assist inner-city entrepreneurs, . . . invest in affordable housing, and . . . help mobilize private lenders" (*Washington Post* 1992). While this plan has been scaled back since then (*Washington Post* 1993c), the Clinton administration clearly believes that community banks can both be successful and help alleviate poverty. Hyman Minsky, et al. (1992: 2) also stress the importance of community banks in eliminating what the authors perceive to be a failure in the market for credit in poor, urban areas. Community development banks, they believe, will

> deliver credit, payment, and savings opportunities to communities not well served by banks and . . . provide financing throughout a designated area for businesses too small to attract the interest of the investment banking and normal commercial banking communities.

The evidence suggests otherwise. Successful community banking projects, both in the United States and abroad, have occurred in communities with relatively homogeneous and geographically immobile borrowers. This has allowed lenders to monitor prospective borrowers to assess their creditworthiness. Monitoring is a substitute for a trouble-free credit history that commercial banks require. In contrast, unsuccessful projects have faced mobile, heterogeneous populations, hindering the ability of lenders to monitor prospective borrowers.

This paper begins by examining existing community banks and related institutions. We highlight the role of information in explaining

---

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

Ralph Chami is Assistant Professor of Economics at the University of Notre Dame. Jeffrey H. Fischer is Deputy Assistant Director in the Bureau of Economics at the Federal Trade Commission. The authors would like to thank Vandana Chandra for helpful comments. The views expressed in this paper do not necessarily reflect those of the Federal Trade Commission or any individual commissioner.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

differences between successful and unsuccessful experiences, and then present a model of borrowing that uses some stylized facts to explain (1) why commercial lenders are reluctant to lend in certain areas and (2) why community lending works in some of those areas but not others. Examining the Clinton proposal in light of this model, along with the work of Minsky et al. (1992), we note that the plan fails to overcome existing informational barriers that make commercial lenders reluctant to operate in certain areas, and is, therefore, unlikely to succeed.

Successful Community Banks

A number of community banking projects have been successful, both in the United States and abroad. In this section we examine some of these projects. We find that the evidence is consistent with our hypothesis that ease of monitoring borrowers, along with a geographically immobile, homogeneous borrowing population are important market characteristics.

The Grameen Bank

This Bangladesh "bank," based in Dhaka, was started in 1976 by Mohammed Yunus in order to give access to capital to the rural poor. The operation targets those who would otherwise be denied credit by the commercial banking sector. Loan amounts are typically small, averaging about $100 in 1985. Repayment rates have averaged about 99 percent within two years of loan disbursement, a higher rate than United States commercial banks typically achieve (World Development Report 1989: 117).

The bank organizes potential borrowers into five-person circles. No collateral is required. Instead, each circle member "must establish a regular pattern of weekly saving before seeking a loan" (ibid.). We refer to this as pre-loan monitoring. In addition, "the first two borrowers must make several weekly payments on their loans before other group members can borrow" (ibid.). Each circle member agrees to repay the loan of any other member, if necessary. If one member defaults on his or her loan, the others are prohibited from borrowing until the loan is repaid (*Wall Street Journal* 1993b). The structure—small circles of borrowers with a mutual interest in seeing loans repaid—gives borrowers the ability and incentive to monitor their fellow circle members.[1] This intra-group monitoring behavior is clearly efficient, and appears again in rotating credit associations, described

[1]For a model in which monitoring can ameliorate moral hazard, see Arnott and Stiglitz (1991).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

below. Commercial banks are typically unable to tie the status of one loan to that of another individual—in the United States this would be a form of redlining and would therefore be illegal—and hence are unable to benefit from the practice. Bank staff also closely monitor field operations, in what can be referred to as post-loan monitoring.

In addition to using monitoring to reduce the rate of default, the bank uses the prospect of future loans to promote repayment. Borrowers know that "the availability of future loans depends on the repayment of borrowed funds" (World Development Report 1989: 117).

Borrowers in rural Bangladesh are unable to take the money and run. For one thing, the small size of each loan reduces the benefits of deliberately defaulting. Second, living in a small village makes disappearing within the village impossible, unlike living within a large city. Third, family ties make relocating to avoid repayment a costlier decision than in a highly mobile society such as the United States. Fourth, the skills residents of rural areas possess are typically such that rural labor is not substitutable for urban labor, again making default less likely. Thus the combination of monitoring and geographic immobility allows the bank to remain profitable despite the bank's inability to screen out high-risk borrowers as do commercial banks.

Finally, circle members tend to be homogeneous, as does the rural population as a whole. Lending to a homogeneous community narrows the distribution of borrowers as measured by the *ex ante* probability that a borrower will repay the loan. Thus screening is less important than in highly heterogeneous communities. This benefit, however, accrues to commercial lenders as well as the Grameen Bank, should a commercial bank choose to lend to such a homogeneous community of borrowers.

Rotating Credit Associations

A second type of informal, community "bank" that has enjoyed success in a number of forms has been the rotating credit association. These associations often share many of the characteristics that made the Grameen Bank successful: group homogeneity, pre- and post-loan monitoring, and, at times, limited mobility outside the group.

Members of a typical rotating credit association contribute a set amount on a regular basis into a common kitty. The use of the funds rotates among the members until each has had a turn. Then the rotation may begin again or the association may disband. The funds are generally used to start small businesses, although Ivan Light and Edna Bonacich (1988) note that, particularly among the Korean community, saving for weddings and funerals is also common.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Rotating credit associations are useful for purchasing objects that cannot be jointly owned, and cost more than a single member can purchase at one time. Timothy Besley, Stephen Coate, and Glenn Loury (1993) show that, for such goods, rotating credit can be an efficient use of capital. All members of the association except the last one can purchase the object earlier than if they did not belong to an association, while the last member purchases the object at the same time with or without the association. For example, suppose each of 10 members needed $50,000 to start a business and each could save $5,000 per year on his own. Without access to capital through a commercial bank or other lender, members must each wait 10 years to start the business. The opportunity cost of the $50,000 saved in each of the intervening years is quite large. By forming a rotating credit association, the individuals can collectively raise $50,000 each year, starting one member in business every year. Capital never accumulates more than a single year before being put to use, and every member but the last starts his business one or more years before he otherwise would.

To make the system work, however, members who have already had use of the funds must be persuaded to continue contributing toward the fund until each member has had a chance to use the capital. Hence family and community ties are particularly important for the success of the association. Besley, Coate, and Loury (1993: 794) describe the group characteristics that make rotating credit associations successful:

> For Roscas [rotating credit associations] to operate successfully it is necessary that individuals keep their commitment to pay into the Rosca after they have won the pot. This may appear problematic since Rosca members are often not able to borrow in conventional credit markets precisely because they cannot be presumed to repay such loans. Roscas circumvent such default problems by exploiting individuals' social connectedness. This is borne out in the anthropological literature, which reveals how the incentive to defect from a Rosca is curbed by social constraints. Roscas are thus typically formed among individuals whose circumstances and characteristics are well known to each other. Defaulters are sanctioned socially as well as being prevented from any further Rosca participation.

Thus social sanctions supplement monitoring to ensure that early users of the fund continue to participate; that is, that the borrowers repay the loan. Having a small group of participants who know one another is particularly useful for this process. Relatives, residents of villages, small towns, and insular communities within larger towns or cities are bound by social contracts, which reduces the risk of default.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

One well-documented form of a rotating credit association is the Korean institution of a *kye*. This practice dates back at least several hundred years and is as much a social institution as a way of raising capital. In the Korean immigrant community in Los Angeles, where capital is difficult to come by since recent immigrants often have no collateral or credit history, the kye has become an important source of capital for small businesses within the community. (Loans from more established immigrants, whether friends or family, also play an important part in this process.)

Members of a kye contribute money, usually weekly, into a common fund which then rotates among the participants. Social pressure to continue payments even after a member has used the funds helps ensure the viability of the kye. Light and Bonacich (1988: 252-53) present a great deal of anecdotal evidence on the prevalence of kyes in the Los Angeles Korean community. They cite an estimate from the *Korean Times* that there were at least 1,000 kyes in Los Angeles, with funds ranging from $500 to $100,000. Many kyes are used for the development of small businesses; one survey the authors cite found that 70 percent of regular kye practitioners were self-employed or planned to be self-employed in the future.

Besides the use of kyes, Korean immigrants rely on a network of more established friends and relatives to raise funds. Light and Bonacich describe a successful gasoline service station owner who learned how to operate a station by working at a relative's station before purchasing his own. In a telephone survey of Korean entrepreneurs, the authors found that 20.7 percent of firms employed relatives or extended kin, averaging 2.4 persons per firm. Employing friends or relatives in an apprentice-like role before lending money allows the lender to engage in a substantial amount of pre-loan monitoring. In addition, familial ties reduce the cost of post-loan monitoring. Unlike commercial banks, which must rely on documentary evidence that a borrower is a good credit risk, local lenders can observe the work habits and even the entrepreneurial skill of a potential borrower and screen out high-risk borrowers. This is particularly useful for recent immigrants, for whom documentary evidence of the sort that commercial banks would require may not be available.

The *Washington Post Magazine* (1993) reports an example, on a smaller scale, of patrons in an Arlington, Virginia bar who have started a fund similar to a rotating credit association. Borrowers must be known to the bar regulars, are limited to small loans, and must explain how they will repay the money. No collateral is required, but default rates nevertheless remain low. As with kyes and other forms of local lending, pre-loan monitoring reduces the riskiness of the loans, and

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

social sanctions (being banished from one's favorite bar) promote repayment.

Besley, Coate, and Loury (1993: 793, fn. 3) note that the 19th-century tradition of a barn raising is also a form of rotating credit association. Instead of contributing money, each member contributes his labor, allowing neighbors to raise barns that otherwise would be too difficult for one family to handle. Community pressure helped ensure that neighbors whose barns were raised early in the process would continue to contribute their labor to subsequent raisings.

South Shore Bank

The institution that appears closest in spirit to the Clinton proposal is Chicago's South Shore Bank. Formed in 1973 in an effort to revitalize the South Shore community, the bank makes loans to residents for community projects. The bank started with home mortgages and small business loans, services that commercial banks were no longer willing to supply.

Lending to community residents allows the bank a greater degree of pre- and post-loan monitoring than a commercial bank would have. Once the loan has been made, for a mortgage on a local house or to start a local business, the borrower is tied to the community, strengthening what were presumably existing ties. Combined with community pressure, these ties help raise the probability of repayment.

The *Washington Post* (1993a: F26) credits the South Shore Bank for revitalizing its community "by doing what a local bank is supposed to—know its neighborhood and borrowers, assist small businesses, lend to homeowners and show a profit." Earlier, the *Post* (1992: A1), quoting Richard Taub of the University of Chicago, found that the bank's success was in part because the "bank found a community that still had people who were working and had pride. You need some strengths to build on."

Monitoring is also important. In a profile of a resident who borrowed from the bank to keep her business going, the *Post* (ibid.) noted that "the bank liked her spirit and her neighborhood reputation." The article continues,

> In its early days, the bank made home mortgages and some business loans—services no longer being offered by anyone else in the neighborhood. At the same time, bank officers and directors talked with tenants, shopkeepers and landlords to find out what the neighborhood needed.

The bank later "sought out and nurtured several dozen of these budding capitalists." Thus monitoring appears to be important to the profitability of the bank. The fact that commercial banks, which cannot

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

monitor borrowers nearly as effectively as a local bank, were unwilling to provide the same services suggests that the expected rate of return to the commercial bank fell below its required level.

Unsuccessful Community Lenders

The previous section described some successful community banks and other lending mechanisms. Examples of unsuccessful community lending projects are harder to come by. Failures, unless in some way spectacular, generally do not attract the notice of the press, and rarely seek media coverage. Nevertheless, there is some evidence that successful community banks are the exception rather than the rule.

Southside Bank and Grand Rapids, Michigan

Inner-city residents of Grand Rapids, Michigan, appear to have much in common with residents of the South Shore community in Chicago. Commercial banks are unwilling to lend into the community, fueling talk of racism. In an effort to revitalize the community, residents are attempting to start the Southside Bank, "a community institution that will focus on financing development" in the area by recycling "deposits from the area into loans to support investment there" (*Wall Street Journal* 1993a: A1).

The project, however, is having difficulty attracting sufficient start-up capital. According to the city's acting community development director, drugs—particularly crack—and crime have hampered the bank's efforts. Another difficulty appears to be the way in which the bank is attempting to raise capital. Community officials have expressed disappointment with commercial banks for their reluctance to fund the Southside Bank. The *Wall Street Journal* (ibid.) says that "the Grand Rapids debate is over whether to promote development by pushing existing banks to lend more in inner-city areas or to establish special community banks."

Yet for commercial banks to support either alternative would be surprising.[2] Commercial banks have already made the business decision that lending into certain communities is not profitable, no doubt in part because the banks are unable to screen attractive candidates from unattractive ones, and in part because the banks have no alterna-

[2]For example, under the Clinton administration, the Civil Rights Division of the Department of Justice has targeted certain banks, such as Chevy Chase Federal Savings Bank in Maryland, for having an insufficient number of branches in poor urban areas. The DOJ's claim is that by operating branches in profitable areas only, banks deny the poor access to inexpensive credit. Residents of poor neighborhoods rely on pawnshops and check-cashing shops to provide loans at interest rates above the rates commercial banks charge.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

tive mechanism for lowering the probability of default, such as monitoring. Simply investing in community banks could be a way for commercial banks to engage in pre- and post-loan monitoring, but at a cost. First, investing in a local bank generates a secondary agency problem. Not only does a commercial bank have to be concerned with actions its borrowers take that might reduce the odds a loan is repaid, but the bank must now be concerned with actions the local bank might take that would be opposed to the investor's interest. Second, commercial banks *do* look to make profitable loans. The bank will take the upper tail of the distribution of loan applicants in any area (though the size of that tail might be small in inner-city communities). These loans are the most likely to be repaid and hence are the most profitable loans. By investing in a local lender, the commercial bank no longer makes these profitable loans itself and receives only a fraction of the profits when the loan is made through the local bank. Other loans are made that, if profitable, will partly offset the loss of those loans the commercial bank was already making, but because the remaining loans are by definition riskier, the marginal profit on these loans is likely to be smaller.[3]

Few Other Community Banks Exist

Other evidence of the difficulty of starting and sustaining a community bank comes more from the absence of such banks than specific examples. The *Washington Post* (1993b) notes that "not many banks . . . have successfully invested in poor communities." Other than the South Shore Bank, there are but two community development banks in the United States: the Self-Help Development Bank, Durham, N.C., and the Community Capital Bank, Brooklyn, N.Y.[4] The *Post* (1993b: C1) explains why,

> Few bankers in the United States have the combination needed—the ability to judge and monitor small and unconventional loans, plus the entrepreneurial bent to encourage businesses and would-be homeowners to expand the investment base in their neighborhoods.

If this explanation is correct, we are unlikely to observe many successful community banks in the United States, and those we do see are likely to be subsidized.

[3]Part of the difference may be made up in the ability of the local bank to better monitor its borrowers and hence reduce the riskiness of any particular loan. Whether the net effect is to raise or lower expected profits to the commercial bank is then unclear.

[4]According to the *Wall Street Journal* (1993a: A1), the number is five, including the South Shore Bank.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Related Institutions

Beyond the traditional types of lending arrangements are other examples of how monitoring plays an important role in extending credit. In this section we discuss two such examples: the German practice of having bank representation on corporate boards of directors, and the inability of sharecroppers to obtain credit from sources other than their landlords.

The Link between Corporations and Banks in Germany

German banks take a more active role in monitoring their corporate borrowers than do United States banks. It is not unusual to find banks having an equity interest in companies with large lines of credit, or to see representatives from the bank on the company's board of directors. According to the *Institutional Investor* (1993: 35),

> Only about 650 of Germany's nearly 3,000 public limit companies are listed on the stock market . . . and the vast majority of these are controlled by families or banks with scant concern for short-term profits.
>
> These dominant shareholders have access to much more complete financial information than do ordinary shareholders reliant on published figures. Dominant shareholders also tend to control the election of the shareholder representatives on supervisory boards. And the power of bank shareholders is strengthened by their additional role as custodians of bearer shares who often act as proxy for other, smaller shareholders. Furthermore, the banks frequently are important lenders—*Hausbanken*—to the companies they partly own. In other words, concludes Global Proxy's Lufkin, "They're controlling the borrower to whom they're lending the money."

One prominent example of this is the Deutsche Bank's 28 percent share in Daimler-Benz. The bank's chairman is also chairman of Daimler's supervisory board. The Deutsche Bank has more than 100 other seats on various corporate boards, including RWE (a Ruhr-based utility company), and Siemens.

Provision of Credit in Interlinked Transactions

A similar market structure is that of credit provision in interlinked transactions. Clive Bell (1988: 797–98) defines an interlinked transaction as one in which the terms of all trades between individuals are jointly determined. For example,

> the landlord who finances his tenants' consumption and working capital; the employer . . . who gives advances to laborers in return for a claim on their time when he needs them or as part of a contract to employ them for a continuous, extended period; and the trader

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

> or commission agent who finances cultivators on the condition that they sell to, or through, him when the crops go to market.

A common arrangement between sharecroppers and landlords is that the tenant borrows money from the landlord to survive during nongrowing seasons, and repays the loan at harvest time. The tenant is unable to obtain credit through other means because of the moral hazard arising from the loan. The tenant's output is sufficiently variable and his effort sufficiently difficult to monitor. Thus, after receiving the loan, effort falls, the probability of default rises, and the loan becomes unprofitable.[5] Sharecropping internalizes much of the moral hazard problem because the landlord need only monitor output, rather than effort. The tenant receives a fixed proportion of output, leading to an optimal provision of effort (though perhaps at the cost of suboptimally shifting risk from the landlord to the tenant).

The Role of Signals in Community Banking

The previous sections described several successful community banks and similar institutions and one unsuccessful project. Certain market characteristics, such as the ability of local lenders to monitor prospective borrowers or otherwise screen out excessively risky loans, appear pervasive in successful projects but are lacking in the Grand Rapids failure.

This section uses a simple model of banking to formalize the importance of monitoring. We use a variant of the model found in Gruben, Neuberger, and Schmidt (1990). Their concern was with the effects of the Community Reinvestment Act on commercial bank lending in poor neighborhoods. They found that while the CRA had the effect of increasing lending in these neighborhoods, while imposing higher costs of investigating the riskiness of potential borrowers than is optimal for the bank. Our focus is on circumstances under which commercial banks find lending in some neighborhoods to be unprofitable while local lenders may find lending in the same neighborhoods to be either profitable or unprofitable.[6]

[5]As a case in point, Jeffrey Fischer and Ralph Chami (1995) demonstrate that the moral hazard problem arising from government transfers in the context of German reunification leads to lower equilibrium effort.

[6]Dwight Jaffee and Thomas Russell (1976), and Joseph Stiglitz and Andrew Weiss (1981) take a different approach. They ask why banks would set an interest rate and a quantity of loans that generates excess demand for credit. Both papers argue that adverse selection makes the market-clearing interest rate unprofitable. This is because low interest rates cause high-risk borrowers to masquerade as low-risk borrowers, inducing the bank to raise its rate; but high interest rates cause low-risk borrowers to opt out of the loan market, leaving only high-risk borrowers, which is again unprofitable for the bank. Our focus, as well as that of William Gruben, Jonathan Neuberger, and Ronald Schmidt (1990), is not

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

We assume there are two types of neighborhoods: poor (P) and rich (R). A commercial bank already lending in neighborhood R must decide whether to open a branch in neighborhood P. Because the lending market in R is competitive, the equilibrium interest rate i is such that the bank earns zero profits in R after accounting for its cost of funds r_d, defaults, and information-gathering costs. Let $C_R(I_R)$ be the cost of gathering information I_R in neighborhood R. Let $p_R(I_R)$ be the probability a loan is repaid, given information I_R. Then competition implies that, in equilibrium,

$$\pi_R = p_R(I_R)\, i - C_R(I_R) - r_d = 0. \tag{1}$$

As I_R increases, p_R and C_R both rise; we assume p_R is concave while C_R is convex in I_R. Thus more information about the neighborhood and prospective borrowers decreases defaults, though at a decreasing rate, while the costs of gathering information increase at an increasing rate.

When considering entry into neighborhood P, the bank is constrained to offering i, but chooses I_P to maximize profits. However, default rates in P are higher than in R because incomes are lower. Given a common level of information I, $p_P(I) < p_R(I)$ for all I. If the bank's information-gathering cost is the same in both neighborhoods, so $C_P(\cdot) = C_R(\cdot) = C(\cdot)$, then it is straightforward to show that

$$\pi_P = p_P(I_P)\, i - C(I_P) - r_d < 0, \tag{2}$$

so the commercial bank will not operate in P.[7]

A local lender, such as a community bank, faces different costs. Its cost of funds, ρ_d, is likely to be higher than r_d; as a small, undiversified lender, it faces greater risks than a commercial bank. On the other hand, its information-gathering costs $\Gamma(I_P)$ are likely to be lower, so $\Gamma(I) < C(I)$ for all I.[8] In such a case the local bank will be profitable if there exists an I_P such that

$$\pi^l = p_P(I_P)\, i - \Gamma(I_P) - \rho_d > 0. \tag{3}$$

---

on credit rationing as such, but rather on how informational asymmetries may cause lenders to withdraw from a particular market completely.

[7]Compare π_P and π_R; i and r_d are common to both expressions. At $I_P = I_R = I$, $\pi_P - \pi_R = i[p_P(I) - p_R(I)] < 0$ by assumption. Suppose $I_P > I_R$. Then p_P rises but $C(I_P)$ rises even faster, so profits are lower still. Suppose $I_P < I_R$. If $\pi_P(I_P) > 0$ for some $I_P < I_R$, then the bank could reduce I_R and earn positive profits at i in neighborhood R, contradicting I_R as an equilibrium value.

[8]One might imagine that i would change as well, since the local lender, unlike the commercial bank, does not lend in neighborhood R and is therefore not compelled to charge i in both neighborhoods.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

If the functions $\Gamma(\cdot)$ and $C(\cdot)$ are sufficiently close to one another, the inequality in (3) will not be satisfied, so the community bank will fail. If, however, $\Gamma(\cdot)$ is sufficiently below $C(\cdot)$, so that the local lender has a significant advantage over commercial banks in acquiring information, then (3) may be satisfied. Homogeneity of community members gives rise to economies of scope in gathering information. Borrowers with ties to the community allow the lender to gather information before committing to a loan, and are more likely to remain in the community after receiving a loan. Hence the same characteristics that appear to be present in successful examples of community banking—community homogeneity and low mobility—are those that are necessary in the model to improve monitoring enough to make local lending profitable.

This argument is echoed in Gruben, Neuberger, and Schmidt (1990), and in Besley, Coate, and Loury (1993). The latter note that ethnic groups and recent immigrants in the United States are more likely to be intimidated by banks and discriminated against and hence less likely to have access to external credit markets. Applicants from those groups often have unreliable income streams, which makes them unable to repay loans on a steady basis, or unwilling to repay, and commercial banks lack effective sanctions. As a consequence, commercial banks perceive the default risk to be too high. Rotating credit associations provide societal sanctions which reduce the default rate, and better *ex ante* information about the reliability of borrowers reduces risks. Besley, Coate, and Loury (1993) describe this as a situation where borrowers offer "social collateral."

Similarly, George Borjas (1992) argues that ethnic differences in skills and earnings are transmitted across generations through investments in human capital and what he terms "neighborhood externalities." That is, both family-specific human capital and "ethnic capital"—the average quality of the ethnic environment—are important factors affecting entrepreneurial skill, work habits, and earning potential. Thus the ability of obtaining a loan may differ across ethnic boundaries. A community bank that is successful in a Korean neighborhood may not be successful in a Hispanic neighborhood because the probability of loan repayment is perceived to be different.

The Clinton Plan Ignores the Conditions Necessary for Success

One of President Clinton's campaign promises was to increase the amount of credit available in poor, urban areas by establishing a system of 100 community banks. While the Clinton administration has scaled

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

back this proposal, it has not disappeared. A variant of the Clinton plan is embodied in the Community Development Banking and Financial Institutions Act of 1993 (HR 3474, approved by the House in November 1993). The act establishes the means to fund community banks in order to increase the amount of capital available in traditionally undercapitalized areas of the country. The bill includes funding of $382 million between fiscal years 1994 and 1997.

Minsky et al. (1992) present a positive view of the Clinton plan. First, the authors argue that there is need for community banks to provide access to credit in poor areas because the private market fails to do so. Second, community banks will be able to provide the services that commercial banks do not, while third, community banks will be profitable, so no ongoing subsidy is needed. HR 3474 also assumes that the banks will be profitable. The act requires that proposals for funds demonstrate that the project will become "an entity that will not be dependent upon assistance from the Fund for continued viability."

Leaving aside the issue of whether there exists a market failure,[9] it is unrealistic to believe that the majority of community banks will be profitable. First, the fact that commercial banks choose not to establish branches in these areas indicates that such branches are not profitable.[10] Minsky et al. (1992) explain this by asserting that commercial banks find the size of loans and deposits in these areas too small to be profitable. This would appear to require an underlying notion of fixed costs of branch-banking, which may be plausible. However, if so, community banks must somehow be able to provide the same services more cheaply than do commercial banks. Since a commercial bank can, presumably, be of any size, the argument relies on the belief that government-funded institutions are somehow more efficient than privately funded institutions. Based on experience with other government-funded institutions, this is unlikely to be the case.

Second, and related to the first point, is that none of these proposals examines *why* commercial banks do not find lending in these areas

[9]Economists typically define a market failure as a situation in which a product is not offered despite the existence of a profitable point on the market's demand curve. Transactions costs may prevent the market from reaching this equilibrium. In the case of banking services, the cost of providing these services presumably exceeds the willingness of the poor to pay for the services. Hence while the "market" for credit in poor areas may not exist, it is not market failure in the traditional sense.

[10]Banking regulators often use the Community Reinvestment Act as a stick with which to force commercial banks to establish branches and engage in some lending in poor areas. Regulatory approval for bank activities, such as mergers, is often dependent on how the bank is perceived to have satisfied the requirements of the Community Reinvestment Act. Commercial banks, however, find these branches to be unprofitable, and view them as a cost of doing business in profitable areas.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

to be profitable. Lenders need some way of assessing the riskiness of a particular loan, and of separating high-risk borrowers from low-risk borrowers. We have argued that situations in which lenders can monitor prospective borrowers can act as a substitute for commercial banks' reliance on credit histories to assess risks. Thus the Grameen Bank, various rotating credit associations, and, to some extent, the South Shore Bank, are all able to achieve low default rates through monitoring.[11] The conditions necessary for a local lender to engage in pre- and post-loan monitoring include a relatively homogeneous community and relative immobile community members.[12] These conditions are often lacking in urban areas of the United States, suggesting that areas commercial banks find unprofitable may also be unprofitable to community banks. As a result, the Clinton plan will either require ongoing subsidies to many community banks or will fail to provide additional capital to poor areas in the long run.

If community banks in fact require ongoing subsidies to maintain their viability, then two policy questions become relevant. First, what are the social gains from subsidizing community banks; and second, is subsidization the most efficient way to achieve those gains? Poor areas that gain access to capital may use much of it to revitalize neighborhoods or to start successful businesses where the social payoff exceeds the private return. Reducing poverty may lower the crime rate, reduce welfare rolls, and decrease health-related problems. If so, costs related to crime prevention, incarceration, unemployment compensation, and Medicaid would fall, and society would benefit. But the cost of community banking, already budgeted for $382 million over the next four years, is substantial, and the ongoing costs are likely to drive the total even higher. Without further study, it is not clear that the benefits of subsidization outweigh the costs.

Turning to the second issue, the most efficient way to encourage neighborhood revitalization may not be to start community banks, but to directly subsidize the desired activities. As with the question of society's gain from community banking, this issue becomes irrelevant if community banks are self-sufficient. If continuing funding is necessary, however, it becomes important to address this issue.

[11]Presumably, commercial banks find using credit histories to be a lower-cost method of screening borrowers than monitoring. In the case of people living in poor areas, or recent immigrants, credit histories are often unavailable or unsatisfactory.

[12]Commercial banks are less concerned with the mobility of their borrowers. Credit reports are nationwide in scope, so a borrower in one area who defaults on a loan cannot simply move to another area and expect to receive future credit.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Conclusion

Government sponsorship of community banks is becoming a reality. The motivation for this policy is apparently threefold: first, that poor communities, particularly in the inner city, have excess demand for capital and other banking services; second, that commercial banks do not find it profitable to provide such services; and third, that it has been proven that community banks and related institutions can be successful in providing capital to these areas.

However, as we have shown, successful community banks, such as the South Shore Bank and the Grameen Bank, and successful rotating credit associations, such as the Korean kyes, occur in communities with special characteristics. Pre-loan monitoring allows lenders to reduce default risks, while post-loan monitoring, close-knit communities, and a relatively immobile population all help ensure that loans will be repaid. Under these circumstances lenders can ameliorate the adverse selection and moral hazard problems inherent in agency relationships under uncertain information and with heterogeneous agents. Monitoring functions as a substitute for the more conventional use of favorable credit histories as a prerequisite for obtaining a loan.

When the Clinton plan is viewed within this framework, community banking in the United States on a large scale is unlikely to be profitable. Inner cities may need capital, but the monitoring mechanism that works so well in rural Bangladesh is not available in urban America. Minsky et al. (1992) recommend community banks as a means of correcting the market "failure" of commercial banks in inner cities, but do not provide a compelling explanation for how the market has failed. Before committing several hundred million dollars for another federal program, we should make certain that we are not merely providing subsidies in disguise.

References

Arnott, R., and Stiglitz, J.E. (1991) "Moral Hazard and Nonmarket Institutions: Dysfunctional Crowding Out or Peer Monitoring?" *American Economic Review* 81 (March): 179–89.

Bell, C. (1988) "Credit Markets and Interlinked Transactions." In Chenery, H., and Srinivasan, T.N. (eds.) *Handbook of Development Economics*, Vol. 1, 763–830. Amsterdam: Elsevier Science Publishers.

Besley, T.; Coate, S.; and Loury, G. (1993) "The Economics of Rotating Savings and Credit Associations." *American Economic Review* 83 (September): 792–810.

Borjas, G.J. (1992) "Ethnic Capital and Intergenerational Mobility." *Quarterly Journal of Economics* 106 (February): 123–50.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Fischer, J.H., and Chami, R. (1995) "Transfers and Incentives: Some Microeconomic Consequences of German Reunification." *Journal of Comparative Economics*, forthcoming.

Gruben, W.C.; Neuberger, J.A.; and Schmidt, R.H. (1990) "Imperfect Information and the Community Reinvestment Act." Federal Reserve Bank of San Francisco *Economic Review* (Summer): 27–46.

Institutional Investor (1993) "Germany Discovers the Foreign Investor," July: 34–40.

Jaffee, D.M., and Russell, T. (1976) "Imperfect Information, Uncertainty, and Credit Rationing." *Quarterly Journal of Economics* 90 (November): 651–66.

Light, I., and Bonacich, E. (1988) *Immigrant Entrepreneurs: Koreans in Los Angeles 1965–1982*. Berkeley: University of California Press.

Minsky, H.P.; Papadimitriou, D.B.; Phillips, R.J.; and Wray, L.R. (1992) "Community Development Banks." The Jerome Levy Lectures in Economics. Bard College Institute, Annandale-on-Hudson, Working Paper No. 83.

Stiglitz, J.E., and Weiss, A. (1981) "Credit Rationing in Markets with Imperfect Information." *American Economic Review* 73 (June): 393–410.

Wall Street Journal (1993a) "Inner-City Capitalists Push to Start a Bank for Their Community." 19 January: A1.

Wall Street Journal (1993b) "Clinton Looks to a Bank in Bangladesh for Model to Help U.S. Poor Get Loans." 27 August: A2.

Washington Post (1992) "Banking on Community Development in Chicago, Clinton Finds a Model to Boost Inner-City Lending." 20 December: A1.

Washington Post (1993a) "Clinton's Risky Plan for Development Banks." 16 January: F26.

Washington Post (1993b) "White House Eyes Community Lending Aid." 15 May: C1.

Washington Post (1993c) "White House Scales Back Community Lending Plan." 19 May: F1.

Washington Post Magazine (1993) "The Loan Arranger." 3 October: 8–9.

World Development Report (1989) Washington, D.C.: The World Bank.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

COMMUNICATIONS

The Drug War and the Homicide Rate: A Direct Correlation?

Harold J. Brumm and Dale O. Cloninger

The nation's concern with drug-control policy has escalated in recent years. This heightened anxiety is reflected in the growth of federal drug-control budget authority appropriated by the Congress. In fiscal year 1985 these budgetary obligations were $2.75 billion; by 1994 they had risen to $12.14 billion (U.S. Department of Justice 1994a: 19).[1]

The out-of-pocket drug-control expenditures made by the criminal justice system—at all levels of government, not just federal—represent only a portion of the total cost of drug enforcement activities. Economists, most notably Milton Friedman (1991: 57) and William Niskanen (1992: 238), have argued that the war on illicit drugs has diverted police resources away from other law enforcement activities, with the result that violent crimes and crimes against property have been higher than they would otherwise have been.[2]

To the extent that communities divert law enforcement resources from violent crimes to illegal drug offenses, the risk of punishment for engaging in violent crimes is reduced. A reduction in this risk would be expected to increase the incidence of violent crimes. This paper seeks to determine empirically the response, if any, in the

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

Harold J. Brumm is Senior Economist at the U.S. General Accounting Office. Dale O. Cloninger is Professor of Finance and Economics at the University of Houston-Clear Lake. The views expressed in this paper are those of the authors and should not be attributed in any way to the U.S. General Accounting Office.

[1]The Department of Justice expresses this budget authority in current dollars. Expressed in constant dollars, this increase would, of course, be somewhat less.

[2]According to the FBI's definition (U.S. Department of Justice 1994b: 10), *violent* crimes are murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. *Property* crimes, according to the FBI's definition (U.S. Department of Justice 1994b: 35), are burglary, larceny-theft (except motor vehicle theft), motor vehicle theft, and arson. For definitions of these individual crimes, see U.S. Department of Justice (1994b: 380).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

homicide offense rate to changes in the percentage of arrests attributed to drug offenses. The empirical results obtained are consistent with a priori expectations that homicide offense rates are higher in communities that devote a greater percentage of their policing resources to the enforcement of drug laws. These results are used to estimate the value-of-life loss due to the war on drugs.

Previous Research

Gary Becker's (1968) seminal work has provided the theoretical underpinnings for the common-sense notion that crime rates will be negatively associated with the opportunity costs of criminal behavior and with the threat of sanctions imposed for criminal offenses (see also Ehrlich 1974, 1975). Samuel Cameron (1988: 308) has provided a stylized version of the econometric model used by empirical researchers to test these and other predictions of the economic theory of crime. In this model the demand for deterrence, the supply of offenses, and the demand for police are jointly determined. Bruce Benson et al. (1992) have used data from Florida's 67 counties for 1986 and 1987 to examine whether property crime is positively related to the intensity of drug enforcement activities. As a proxy for the intensity of these activities, they used the number of drug arrests as a proportion of the number of index-I arrests.[3] Their principal finding is that a 1-percent increase in drug enforcement activities raises the property crime offense rate by an estimated 0.164 percent (Benson et al. 1992: 689).

To generate the empirical results presented below, we use a two-equation variant of the standard three-equation econometric model. The equation dropped refers to the demand for police; we treat police quantity demanded as exogenous. William Trumball (1989: 428), and Helen Tauchen, Anne Witte, and Harriet Griesinger (1993: 11) report evidence in support of this exogeneity. For the present paper, the argument for exogeneity is derived from the fact that the data used for this study show the distribution of police resources to be dependent on crimes other than homicide (Cloninger 1992: 416).

[3]Benson et al. (1992: 682) define index-I crimes as murder, forcible sexual offenses, aggravated assault, robbery, burglary, larceny, and auto theft. This taxonomy is close, if not identical, to what the FBI calls "Part I" crimes. Offenses in the FBI's Uniform Crime Reporting are divided into two groupings, Part-I and Part-II (U.S. Department of Justice 1994b: 380). The Part-I offenses are: criminal homicide (murder and nonnegligent manslaughter, and manslaughter by negligence), forcible rape, robbery, aggravated assault, burglary, larceny-theft (except motor vehicle theft), motor vehicle theft, and arson.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

The Empirical Model

The data used to estimate the model are from 59 cities in 32 states, and reflect 1985 experience (Cloninger 1992: 416).[4] The year 1985 was chosen because of data availability considerations. Two of the cities were dropped from the sample because of missing data. Thus, each observation in the data base used for the present study is a city, and the sample size is 57.

The model's variables, which are expressed as natural logarithms, consist of two that are endogenous, and five that are exogenous. The former are the homicide offense rate (LHRATE), and the homicide arrest rate (LHARST); the latter are the percentage of the population that is nonwhite (LNONW), the number of police per violent offense (LCPVO), the number of residents per square mile (LPOPDENS), the median value of owner-occupied housing (LHOUSE), and the number of drug arrests as a percentage of total arrests (LDRUGPRO). The last variable is a proxy for the war on drugs.[5]

The model's equations are

$$(1) \qquad LHARST = \beta_{10} + \beta_{11}LHRATE + \beta_{12}LCPVO + \beta_{13}LPOPDENS + \beta_{14}LDRUGPRO + \epsilon_1,$$

$$(2) \qquad LHRATE = \beta_{20} + \beta_{21}LHARST + \beta_{22}LNONW + \beta_{23}LHOUSE + \epsilon_2.$$

The errors in equations, ϵ_1 and ϵ_2, are assumed to be normally distributed with a mean of zero and a constant variance.

The two hypotheses of greatest interest here are that an increase in drug arrests relative to total arrests will reduce the homicide arrest rate, that is, $\beta_{14} < 0$, and that the latter reduction will—according to the economic theory of crime—lead to an increase in the homicide offense rate, that is, $\beta_{21} < 0$. The economic theory of crime makes two additional predictions. One is that an increase in the opportunity cost of crime constrains criminal behavior. LHOUSE is a proxy for this opportunity cost. Thus, the theory predicts that $\beta_{23} < 0$. The other prediction is that an increase in police resources will increase criminal justice system output, implying that $\beta_{12} > 0$.

[4]The data were obtained from four sources: U.S. Department of Commerce (1986); U.S. Department of Justice (1986); Cohn and Sherman (1986); and private communication with J. Harper Wilson, Chief, Uniform Crime Reporting Section, Federal Bureau of Investigation, U.S. Department of Justice, Washington, D.C., July 22, 1987.

[5]An anonymous referee has pointed out that LCPVO and LDRUGPRO are not strictly exogenous. That said, in the sample used for this study homicide commissions are an extremely small percentage of the total number of crimes committed. For this reason, and because of data limitations, LDRUGPRO is treated as exogenous. The exogeneity of LCPVO was justified in the previous section.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Regarding the sign of β_{11}, there are two competing hypotheses. The resource-saturation hypothesis (Fisher and Nagin 1978: 364) predicts that $\beta_{11} < 0$; the toughening hypothesis (Fisher and Nagin 1978: 365) predicts that $\beta_{11} > 0$. The first hypothesis asserts that as the number of crimes committed goes up, the number of arrests made will—for a police force of a given size—increase but at a decreasing rate. At the margin, this will reduce the risk of punishment for committing a crime. The second hypothesis argues that the criminal justice system will respond to increased offense rates by toughening sanctions, which will at the margin increase the risk of punishment. The resource-saturation hypothesis is consistent with the view that increased drug enforcement activities divert scarce policing resources from controlling other offenses, thereby reducing the risk of punishment for committing those offenses.

The variables LPOPDENS and LNONW were included in the model to control for the potential impacts of phenomena thought by sociologists and criminologists (e.g., Gottfredson and Hirschi 1995) to be important in explaining criminal arrest and offense rates. The economic rationale, according to Isaac Ehrlich (1974: 127; 1975: 412), for including LNONW in a criminal offense supply function is that nonwhites have a relatively low opportunity cost of crime commission. If so, then economic theory predicts that $\beta_{22} > 0$. But the theory is silent about the sign of β_{13}. However, it seems reasonable to anticipate that $\beta_{13} < 0$, that is, that the more densely populated a city is, the more difficult for the police to ferret out offenders.

The joint determination of LHARST and LHRATE requires the use of an estimation method which accounts simultaneously for that interdependency. System estimation techniques, such as the method of three-stage least squares (3SLS), yield smaller estimated coefficient standard errors than do single-equation estimation techniques, such as the method of two-stage least squares (2SLS). However, the individual parameter estimates obtained by system estimation techniques are relatively more sensitive to the specification of the entire model—a specification error in one equation can adversely affect the parameter estimates in all of the model's equations. That is, there is a tradeoff between potential costs of specification error and a gain in statistical efficiency if a system estimation method rather than a single-equation estimation method is chosen (Pindyck and Rubinfeld 1981: 339). This tradeoff was avoided by using both 2SLS and 3SLS to estimate equations (1) and (2).[6]

[6]MicroTSP (Hall, Johnston, and Lilien 1990) was the software package used to generate the estimates.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Tables 1A and 1B present the results of 2SLS estimation of the model. The signs of all estimated coefficients are those that were anticipated. The estimated standard errors, t-statistics, and p-values indicate that, except for the intercept in the deterrence-demand equation, all estimated coefficients are statistically significant, albeit weakly in some cases. Each estimated p-value is at a level at which the observed t-statistic would just be significant (Goldberger 1991: 239–40).

TABLE 1A

DEMAND FOR DETERRENCE
(2SLS METHOD)

Explanatory Variables	Coefficient	Standard Error	t-Statistic	(p-value)
INTERCEPT	−1.0976	0.7207	−1.5230	0.1338
LHRATE	−0.1093	0.0600	−1.8221	0.0742
LCPVO	0.1136	0.0644	1.7646	0.0835
LPOPDENS	−0.0894	0.0346	−2.5842	0.0126
LDRUGPRO	−0.0699	0.0367	−1.9043	0.0624
Adjusted R-squared	0.3299			
Number of observations	57			

TABLE 1B

SUPPLY OF HOMICIDES
(2SLS METHOD)

Explanatory Variables	Coefficient	Standard Error	t-Statistic	(p-value)
INTERCEPT	6.5851	1.9039	3.4588	0.0011
LHARST	−2.4562	0.7768	−3.1621	0.0026
LNONW	0.2700	0.1542	1.7512	0.0857
LHOUSE	−0.5061	0.1839	−2.7519	0.0081
Adjusted R-squared	0.1211			
Number of observations	57			

For example, if the null hypothesis that the coefficient of LDRUGPRO equals zero were true (H_0: $\beta_{14} = 0$), then the probability of obtaining −0.0699 as an estimate of that coefficient (and, hence,

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

of obtaining −1.9043 as the estimate of the corresponding *t*-statistic) is equal to just .0624. Finally, note that the negative sign for the estimated coefficient for LHRATE in the deterrence-demand equation provides support for the resource-saturation hypothesis.

Tables 2A and 2B present the results of 3SLS estimation of the model. Again the signs of all estimated coefficients are those that were anticipated. Not surprisingly, all 3SLS estimated coefficient standard errors are smaller than their 2SLS counterparts.

TABLE 2A

DEMAND FOR DETERRENCE
(3SLS METHOD)

Explanatory Variables	Coefficient	Standard Error	*t*-Statistic	(*p*-value)
INTERCEPT	−1.4639	0.6666	−2.1962	0.0326
LHRATE	−0.1295	0.0565	−2.2904	0.0261
LCPVO	0.1547	0.0573	2.7000	0.0093
LPOPDENS	−0.0854	0.0304	−2.8115	0.0069
LDRUGPRO	−0.0441	0.0251	−1.7587	0.0845
Adjusted *R*-squared	0.2246			
Number of observations	57			

TABLE 2B

SUPPLY OF HOMICIDES
(3SLS METHOD)

Explanatory Variables	Coefficient	Standard Error	*t*-Statistic	(*p*-value)
INTERCEPT	5.7915	1.5356	3.7716	0.0004
LHARST	−2.3702	0.7410	−3.1984	0.0023
LNONW	0.3253	0.1312	2.4801	0.0163
LHOUSE	−0.4443	0.1591	−2.7926	0.0073
Adjusted *R*-squared	0.1372			
Number of observations	57			

The elasticity of the homicide offense rate with respect to drug enforcement activities is $(\partial \text{ LHARST}/\partial \text{ LDRUGPRO})(\partial \text{ LHRATE}/\partial \text{ LHARST}) = \beta_{14}\beta_{21}$. The 2SLS estimate of this elasticity is approximately

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

TABLE 3

VALUE-OF-LIFE COST

Estimation Methods	Cost per Capita (1985 $)	Cost per 100,000 Inhabitants (Millions of 1985 $)	1985 Homicide Offense Rate per 100,000 Inhabitants	Actual Cost (Millions of 1985 $)	Elasticity	Estimated Increase in Cost (Millions of 1985 $)
2SLS	2,600,000	260,000	7.9	2,054,000	.00170	3,492
3SLS	2,600,000	260,000	7.9	2,054,000	.00105	2,157

SOURCES: Cohen (1990: 140) for data on value-of-life cost per capita; U.S. Department of Justice (1986: 7) for data on 1985 homicide offense rate.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

0.17 percent.[7] That is, a 1 percent increase in drug enforcement activities is estimated to result in a 17 one-hundredths of 1 percent increase in the homicide rate. For 3SLS, the estimate of this elasticity is approximately 0.105 percent.

The 2SLS and 3SLS estimates of the drug enforcement elasticity of the homicide rate can be converted into estimates of the value-of-life loss due to the war on illicit drugs. This can be accomplished by using value-of-life estimates that have been reported in various studies of labor-market risk premiums (Viscusi 1993: 1926–27). A representative estimate, developed by Kip Viscusi (1983: 106), and updated by Mark Cohen (1990: 140) to 1985 dollars, is $2.3 million. As computed in Table 3, the 2SLS estimate of the increased value-of-life cost that would result from a 1-percent increase in the intensity of drug enforcement activities is about $3.5 billion annually in 1985 dollars.

The corresponding 3SLS estimate is roughly two-thirds of that amount. If the rate of increase in the CPI-U is used as the measure of the inflation rate, then in 1994 dollars the 2SLS and 3SLS estimates are $4.81 billion and $2.97 billion, respectively.

Conclusion

The foregoing estimates of the increased value-of-life cost of the war on illicit drugs are, of course, quite crude. Econometric studies, including the present one, should be taken with at least one grain of salt. However, if the empirical results presented here are even approximately correct, then it is clear that current drug-control policy is substantially more expensive than indicated by the observed out-of-pocket drug-control expenditures made by the criminal justice system.

References

Becker, G.S. (1968) "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 78: 169–217.

Benson, B.L.; Kim, I.; Rasmussen, D.W.; and Zuehkle, T.W. (1992) "Is Property Crime Caused by Drug Use or by Drug Enforcement Policy?" *Applied Economics* 24: 679–92.

Cameron, S. (1988) "The Economics of Crime Deterrence: A Survey of Theory and Evidence." *Kyklos* 41: 301–23.

Cloninger, D.O. (1992) "Intercity Variations in the Police Use of Lethal Response." *Journal of Economic Behavior and Organization* 50: 413–22.

Cohen, M.A. (1990) "A Note on the Cost of Crime to Victims." *Urban Studies* 27: 139–46.

[7]Since the estimates of β_{14} and β_{21} are statistically significant, the estimated elasticity is statistically significant.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Cohn, E.G., and Sherman, L.W. (1986) "Citizens Killed by Big-City Police." Washington, D.C.: Crime Control Institute and Crime Control Corporation.
Ehrlich, I. (1974) "Participation in Illegal Activities: An Economic Analysis." In Becker, G.S., and Landes, W.M. (eds.) *Essays in the Economics of Crime and Punishment*, 68–134. New York: Columbia University Press.
Ehrlich, I. (1975) "The Deterrent Effect of Capital Punishment: A Question of Life or Death." *American Economic Review* 65: 397–417.
Fisher, F.M., and Nagin D. (1978) "On the Feasibility of Identifying the Crime Function in a Simultaneous Model of Crime Rates and Sanction Levels." In Blumstein, A.; Cohen, J.; and Nagin, D. (eds.) *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, 361–99. Washington, D.C.: National Academy of Sciences.
Friedman, M. (1991) "The War We Are Losing." In Krauss, M.B., and Lazear, E.P. (eds.) *Searching for Alternatives: Drug-Control Policy in the United States*. Stanford, Calif.: Hoover Institution Press.
Goldberger, A.S. (1991) *A Course in Econometrics*. Cambridge, Mass.: Harvard University Press.
Gottfredson, M.R., and Hirschi, T. (1995) "National Crime Control Policies." *Society* 32: 30–36.
Hall, R.E.; Johnston, J.; and Lilien, D.M. (1990) *MicroTSP User's Manual, Version 7.0*. Irvine, Calif.: Quantitative Micro Software.
Niskanen, W.A. (1992) "Economists and Drug Policy." *Carnegie-Rochester Conference Series on Public Policy* 36: 223–48.
Pindyck R.S., and Rubinfeld D.L. (1981) *Econometric Models and Economic Forecasts*, 2nd ed. New York: McGraw-Hill.
Tauchen, H.; Witte, A.D.; and Griesinger, H. (1993) "Crime Deterrence: Revisiting the Issue with a Birth Cohort." NBER Working Paper No. 4277. Cambridge, Mass.
Trumball, W.N. (1989) "Estimations of the Economic Model of Crime Using Aggregate and Individual Level Data." *Southern Economic Journal* 56: 423–39.
U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census (1986) *State and Municipal Data Book, 1986*. Washington, D.C.: U.S. Government Printing Office.
U.S. Department of Justice, Federal Bureau of Investigation, Uniform Crime Reporting Section (1986) *Crime in the United States, 1985*. Washington, D.C.: U.S. Government Printing Office.
U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics (1994a) *Sourcebook of Criminal Justice Statistics, 1993*. Washington, D.C.: U.S. Government Printing Office.
U.S. Department of Justice, Federal Bureau of Investigation, Uniform Crime Reporting Section (1994b) *Crime in the United States, 1993*. Washington, D.C.: U.S. Government Printing Office.
Viscusi, W.K. (1983) *Risk by Choice: Regulating Health and Safety in the Workplace*. Cambridge, Mass.: Harvard University Press.
Viscusi, W.K. (1993) "The Value of Risks to Life and Health." *Journal of Economic Literature* 31: 1912–46.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Scholarships: Need or Merit?

Herschel Grossman

The Ivy-MIT Cartel

About 40 years ago the eight colleges in the Ivy group—Brown, Columbia, Cornell, Dartmouth, Harvard, Pennsylvania, Princeton, and Yale—together with MIT formed a cartel to limit competition for desirable undergraduate students. The members of the cartel agreed not to award "merit" scholarships to undergraduates and to give undergraduate financial aid—that is, to offer undergraduates discounts off their full tuition charges—only on the basis of "need."

In this context, merit is shorthand for any characteristic that makes a student especially desirable to a college. Colleges are eager to attract students who have exceptional academic, artistic, or athletic talent not only because most faculty prefer to teach students who are smart and interesting, but also because the academic, artistic, and athletic achievements of talented students enhance the reputation of a college and the value of its degrees. This effect pleases alumni, who are the main benefactors of private colleges, and also makes a college more attractive to other prospective students. In this regard, colleges are like any business for whom "quality" customers enhance the reputation of the product and attract other customers.

By agreeing to give only need-based undergraduate financial aid and by also adopting a niggardly definition of need, the Ivy colleges and MIT were able to enroll desirable students while not giving them more than the minimum necessary amount of financial aid. Without an agreement not to give merit scholarships, competition for desirable undergraduates could have led to large increases in financial aid. Moreover, the Ivy colleges with the largest endowments (Harvard, Yale, and Princeton) probably would have been the most aggressive in using merit-based financial aid to compete for the most desirable

---

Cato Journal, Vol. 14, No. 3 (Winter 1995). Copyright © Cato Institute. All rights reserved.

The author is Merton P. Stoltz Professor in the Social Sciences and Professor of Economics at Brown University.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

undergraduates. But, much of this competition would have been with each other, because Harvard, Yale, and Princeton together attract without merit scholarships a major share of those undergraduates who would have been likely to receive merit scholarships. In forming their cartel the Ivy colleges and MIT apparently recognized that allowing merit scholarships would have depressed their net tuition revenues with little change in the actual composition of their enrollments.

To enforce their agreement to give undergraduate financial aid only on the basis of need, as determined by their formulas, the Ivy colleges and MIT for many years held an annual meeting, called Overlap, at which they decided on the dollar amounts of financial aid to offer to individual prospective students. Over the years, several other private colleges joined the Ivy colleges and MIT as additional participants in Overlap. But, all colleges who competed with the Ivies and MIT for desirable students, whether or not they participated in Overlap, benefited from competition restraint by the Ivies and MIT.

The Ivy colleges and MIT not only have given undergraduate financial aid only on the basis of need, but, with the exception of Brown, they also claim to practice need-blind undergraduate admissions and to provide sufficient financial aid to meet the full need of all students. In other words, their announced policy has been to make admissions decisions without taking into account whether an applicant needs financial aid and then to give each admitted applicant any financial aid that, according to their formulas, the applicant needs.

Penalizing Prudence

This policy of basing undergraduate financial aid only on need, especially when supplemented by need-blind admissions, has the undesirable effect of reducing the incentive for families to make their own provisions for paying college tuition. The larger a family's income and savings and the fewer children it has, the smaller the amount of need-based financial aid for which it qualifies. In this way, need-based financial aid penalizes families who have accumulated sufficient savings to be able to pay to send their children to college. Need-blind admissions goes further and denies such families any advantage in being able to get their children admitted to the college of their choice.

Many economists would argue that because it penalizes prudence, need-based financial aid encourages sensible families to be imprudent. In a recent study, Martin Feldstein (1995) explains how need-based financial-aid formulas act like a tax on the savings of middle-income families. Using data from the 1986 Survey of Consumer Finances, Feldstein estimates that this implicit tax annually causes the typical

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

middle-income family with two precollege children to reduce the amount it saves by about 50 percent. According to this analysis, it is not surprising that the savings of most middle-income families are now inadequate for them to pay to send their children to the Ivy colleges and MIT and that applicants from middle-income families need more and more financial aid.

Need-based financial aid also limits the economic opportunities of women who have left paid employment in order to stay home with their children. Many mothers look forward to resuming their careers when their children go off to college. But, any contribution that mothers of college students make to their family income causes a reduction in need-based financial aid.

If talented undergraduates were able to receive merit scholarships, then the implicit tax on savings and family income would be mitigated. With some undergraduate financial-aid awards based on merit, financial aid would be more than a reward for being poor, as it is now at colleges that give only need-based financial aid. Rather, financial aid would be a prize available to any talented applicant who, on the basis of ability, industry, and any other relevant characteristics, ranks at the top of the applicant pool.

Not for Profit?

The Ivy colleges and MIT defend their prohibition of merit scholarships by claiming that this policy enhances equality of opportunity. Colleges argue that they cannot afford to increase the total amount of undergraduate financial aid that they already give. Consequently, in order to give substantial merit-based financial aid to undergraduates, they would have to decrease financial aid to needy students.

This argument is hardly convincing. In fact, MIT and each of the Ivies have budget options that would enable them to avoid reductions in need-based financial aid even if they awarded merit scholarships to many undergraduates.

The richest Ivy colleges (Harvard, Yale, and Princeton) have substantially larger incomes from endowments and annual gifts than the other Ivy colleges and MIT. The richest colleges, however, devote to financial aid for undergraduates a smaller fraction of their gross revenues from endowment, gifts, and tuition than do many of the poorer colleges. This anomaly is consistent with the view that collusion enabled the richer participants to give less financial aid than they would give under competition. It also suggests that the richest colleges have not been budgeting for undergraduate financial aid the maximum amount that they can afford. The richest colleges could increase

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

significantly the amount of financial aid that they give to undergraduates simply by devoting to undergraduate financial aid the same fraction of their gross revenues as do many of the poorer colleges.

In fact, some colleges who did not participate in Overlap and who are not as rich as the richest Ivy colleges seem to be more generous in giving both need-based financial aid and merit scholarships. For example, according to *The Boston Globe* (see Dembner 1995), Brandeis University annually awards $2.5 million in merit scholarships while also giving $6 million in need-based financial aid, which is sufficient to meet the full need of all students.

Another important issue is the nature of the so-called costs of providing an undergraduate education at MIT and the Ivy colleges. MIT and the Ivy colleges want us to believe that these costs cannot be reduced in order to free up funds for additional financial aid.

One reason to doubt this claim is that, because the tax laws allow private colleges to designate themselves to be not-for-profit enterprises, the costs that colleges calculate are artificially inflated. The not-for-profit designation is one of the methods by which the taxpayers subsidize education. It allows colleges to keep their books in such a way that they report no profits for tax purposes. But this bookkeeping contrivance does not mean that colleges do not earn what an economist would properly call profits. Rather, the not-for-profit designation allows private colleges, and other not-for-profit enterprises, to include in what they call their costs some amounts that are really profits.

In ruling in an antitrust case against MIT, *U.S. v. Brown University, et al.* (1993), a Federal Appeals Court recognized that financial aid is not "charity," but rather "part of the commercial process of setting tuition." As the Federal Appeals Court put it, "The higher than competitive tuition prices which MIT and the other Overlap members were able to charge, absent competition, enhances 'revenues,' if not 'profits,' which can be allocated to any conceivable internal institutional purpose."[1]

Clearly, one such institutional purpose is the generous compensation and perquisites that the faculty and administrators at the Ivy colleges and MIT receive. Those perquisites include light teaching loads, substantial support for research projects and graduate students, and free time to earn consulting income. The costs attributed to

[1]In a recent paper, Dennis Carlton, Gustavo Bamberger, and Roy Epstein (1995) claim that the Overlap cartel did *not* result in higher tuition revenues. In my view, their own evidence, especially their finding that tuition revenues were independent of endowments, actually supports the opposite conclusion, which is the common-sense conclusion that the Appeals Court reached.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

generous compensation and perquisites for faculty and administrators are to some extent really profits.

The correct economic view of private colleges would seem to be that they function similarly to producer cooperatives or partnerships, and that the partners, who in this case are the faculty and administrators, share the profits, which consist of any excess of revenues over true costs. That private colleges attempt to maximize profits is not itself a problem. In fact, the profit motive has made the elite American colleges into great educational institutions. All studies show that, despite collusive restrictions on financial aid, an education at an elite private college is a good investment. But, consumers get maximum value only if the profit motive encounters unrestricted price competition.

Another reason to doubt the claim that the costs of providing an undergraduate education at the Ivy colleges and MIT cannot be reduced is that at least part of the steep increase in costs in recent years probably reflects nonprice competition for desirable students. Overlap precluded the use of merit scholarships to attract desirable undergraduates, but it did not prevent costly recruiting efforts such as visits by admissions officers and athletic coaches to the high schools and homes of prospective students, and paid visits to campus by prospective students. Even more importantly, the prohibition on merit scholarships for undergraduates induced the participants in Overlap to engage in quality competition in order to attract desirable students. This quality competition included the provision of extensive counseling and psychological support systems and the building of lavish facilities for artistic and athletic activities, all of which are peripheral to academic functions. Like any restriction on price competition, the prohibition on merit scholarships caused costs to increase and probably forced students to pay for more expensive services than a competitive market, responsive to customer preferences, would have supported.[2]

In sum, it seems clear that the Ivy colleges and MIT could continue generous need-based financial aid, including need-blind admissions with full need met, even if they also were offering merit scholarships. Of course, they might not do so. Instead, they might reduce need-based financial aid as competition for desirable students intensified and merit scholarships depressed net tuition revenue. But, this outcome, were it to occur, would not be forced on either the Ivies or MIT. Rather, it would result from their choice to reduce need-based financial aid rather than to make alternative budget cuts, including adjustments in the compensation and perquisites of faculty and administrators.

[2]See Grossman (1995) for more discussion of inflated costs at the Ivy colleges, especially with regard to need-based financial aid for recruited athletes.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Educational Opportunity

For decades the Overlap cartel functioned openly and without legal challenge. But, early in 1991, the Justice Department brought an antitrust action (*U.S. v. Brown University, et al.*) against the eight Ivy colleges and MIT for colluding to fix tuition charges to individual students. The eight Ivy colleges quickly entered into a consent decree with the Justice Department in which they agreed to terminate Overlap. (MIT did not join this initial settlement, but after some litigation MIT and the Justice Department reached a separate settlement.)

Since the termination of Overlap, although the Ivy colleges and MIT still claim to be giving undergraduate financial aid only on the basis of need, they seem in practice to be adjusting financial aid offers according to merit, mainly by stretching their formulas for determining need, in order to compete for desirable students. The likelihood is that, if an enforceable prohibition on merit scholarships were not reinstated, one or another of the Ivy colleges or MIT would become incrementally more aggressive in using merit-based financial aid to attract desirable undergraduates. In response, the other colleges, all of whom want to attract the most desirable students, would be forced to match the merit-based financial-aid offers of their competitors. In other words, without an enforceable prohibition on merit scholarships it probably would not be possible for any former participant in Overlap, however committed it is to giving undergraduate financial aid only on the basis of need, to maintain such a policy indefinitely without suffering a noticeable decrease in the talent and diversity of its undergraduates. Consequently, even without either MIT or any of the Ivies being overtly aggressive in awarding merit scholarships, the competition for desirable undergraduates over time probably would intensify until eventually all of the Ivies and MIT would be giving a substantial amount of financial aid to undergraduates on the basis of merit and without regard for need.

Most importantly, this competition for desirable students undoubtedly would force the Ivy colleges and MIT to increase total financial aid to undergraduates. In a competitive equilibrium, more undergraduates would receive financial aid and financial aid packages would be larger. Merit-based financial aid not only would help talented students from middle-income families who do not qualify for need-based financial aid, but also would mean more generous financial aid for talented applicants from poor families.

This outcome, unfortunately, now seems unlikely. In October 1994 the Congress, in the face on intense lobbying by the educational establishment, passed legislation (Public Law 103-382, Section 568) that explicitly permits colleges to agree to give only need-based finan-

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

cial aid, to adopt a common definition of need, and to exchange any information about the income and assets of prospective students and their families that is necessary to make such an agreement work.

This new legislation makes a mockery of the claim that a prohibition of merit scholarships ensures generous financial aid for needy students. Colleges that agree under this legislation to give only need-based financial aid have to practice need-blind admissions, but the legislation does not require them to provide sufficient financial aid to meet the full need of all students.

In effect, all colleges, including the Ivies and MIT, are now legally free to collude to limit financial aid in any way that they choose. By allowing colleges to deny undergraduates the benefits of merit scholarships, the Congress has opted to protect the economic interests of elite institutions of higher education. In so doing, the Congress has sanctioned a policy that subsidizes imprudence and limits educational opportunity.

References

Carlton, D.; Bamberger, G.; and Epstein, R. (1995) "Antitrust and Higher Education: Was There a Conspiracy to Restrict Financial Aid?" *RAND Journal of Economics* 26(1): 131–47.

Dembner, A. (1995) "Colleges Vying for Students Add Merit Aid." *The Boston Globe*, 12 September.

Feldstein, M. (1995) "College Scholarship Rules and Private Saving." *American Economic Review* 85(3): 552–66.

Grossman, H. (1995) "The Economics and Politics of Scholarships." *Academic Questions* 8(3): 59–66.

U.S. v. Brown University, et al., 5 F.3d 658 (3rd Cir. 1993).

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

BOOK REVIEWS

Out of Work: Unemployment and Government in Twentieth-Century America
Richard K. Vedder and Lowell E. Gallaway
Oakland, Calif.: The Independent Institute, 1993, 336 pp.

This is an impressive and ambitious work that debunks powerful myths about unemployment and calls into question conventional wisdom about how labor markets operate. In 15 chapters (plus a technical appendix and an extensive bibliography), the authors apply economic theory and empirical techniques to explain fluctuations in unemployment during the entire 20th century. They convincingly show that unemployment can be explained by movements in real wage rates after taking into account the quality and productivity of labor. Put simply, the laws of supply and demand remain valid for labor markets.

In the first three chapters, the authors develop and explain their model of unemployment, which is so disarmingly simple that its ability to predict the unemployment rate over even a short period of time would be suspect. Yet, amazingly, their model is incredibly accurate in predicting unemployment rates over the entire nine decades from 1900 to 1990 (see Figure 3.1, p. 31). During this protracted period, the labor market was buffeted by financial panics, two world wars, the Great Depression, and major technological, political, and social changes both at home and abroad.

Chapters 4 through 12 contain insightful and interesting discussions of these significant periods in U.S. economic history. Of note is how Vedder and Gallaway trace the formation of the conventional wisdom among economists and members of the business community in some of these major economic upheavals. The authors' examination of the intellectual history of state intervention reveals that economists are influenced by "group think" and self-interest.

Vedder and Gallaway are accomplished writers who transform material that could be esoteric and dull into fascinating reading, even for professional economists. Sadly, all too many recently trained economists have an appalling lack of historical perspective because their "education" has emphasized statistical and mathematical legerdemain rather than the lessons of history, which are often much more difficult to analyze and interpret. Learning from the past is decidedly out of fashion. Both professionals and interested laypersons will gain invaluable insights from *Out of Work*—a book that is urgently needed, long overdue, and cogently written.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Policymakers are well advised to take to heart the message Vedder and Gallaway convey in the last three—and most important—chapters, which investigate the effects of government involvement in labor markets. One sentence encapsulates their findings: "Government actions, as expressed in a wide range of public-policy actions, generally have contributed to the presence of higher levels of unemployment, rather than lower" (p. 264). Since the 1930s, virtually every aspect of the employer-employee relationship has been regulated by federal, state, and local government mandates. Minimum wages, mandated benefits, regulation of the work hours, unemployment compensation, and the promotion of collective bargaining are some of the myriad ways in which politicians have meddled in the labor market over the years.

Who bears most of the costs of this state intervention? In chapter 14, the authors persuasively argue that unemployment differentials have increased markedly during the last three decades for females, minorities, and teenagers—precisely the groups that public policy supposedly is designed to assist. The lesson is simple: Good intentions do not guarantee good public-policy outcomes. Markets respond in predictable ways to political meddling and, always and everywhere, those on the bottom of the economic ladder bear a disproportionate share of the burden.

For at least the last half-century, the oft-repeated mantra of politicians of every stripe in the United States and elsewhere in the world has been "Jobs!" The record high levels of unemployment during the Great Depression left serious, and as yet unhealed, scars on the psyche of Americans, and the Keynesian aftermath, which dictated that government "do" something to relieve unemployment and the associated dislocations, forever changed the economic landscape. Since the 1960s, the government constantly has been involved in the labor market, purportedly to create new jobs or to train potential employees so that their skills were better suited to existing employment opportunities. Over the decades, an alphabet soup of programs and agencies have spent hundreds of billions of taxpayers' dollars, most of which have been squandered on ill-fated and scandal-ridden schemes, such as CETA (Comprehensive Employment and Training Act).

On a common-sense level (and as indicated by the Vedder-Gallaway model), it is patently evident that government cannot possibly "create" jobs; if it could, unemployment and the often-associated ills of poverty would have disappeared long ago. When economic decisionmaking moves from markets to the political arena, political clout prevails—jobs are created for powerful union workers at the expense of other, less fortunate workers. Jobs created by government programs in one sector of the economy inevitably reduce employment opportunities elsewhere; employment is merely shifted around to favor certain groups. Politicians and voters who would never presume that the laws of physical nature could be repealed by legislative fiat are evidently too persuaded by self-interest—or in the case of voters, too gullible—when it comes to attempting to repeal the laws of economics in the labor markets.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Common sense, however, is no longer in vogue among either economists or (not surprisingly) politicians. Thus, this work performs the very significant function of showing both empirically and theoretically that the basic laws of supply and demand still operate in the labor markets, despite the efforts of politicians to repeal those laws.

Let us fervently hope that *Out of Work* is widely read and that its powerful message is heeded: Government has done far more harm than good by meddling in labor markets and has wasted untold taxpayers' money in the process.

James T. Bennett
George Mason University

Federalism and Free Trade
Jean-Luc Migué
London: Institute of Economic Affairs, 1993, 77 pp.

When discussing politics and free trade, economists commonly argue that we must control the special-interest tendencies of government to fully secure the advantages of international trade. While recognizing the importance of this argument, Jean-Luc Migué persuasively points out that the opposite is also true: free trade is needed if we are to secure the advantages of controlling the special-interest tendencies of government. Migué's position is clearly stated in his introduction:

> Provided there is free inward and outward movement of factors and goods, a national government has little power to engage in redistributive policies. Stated conversely, it is the ability to restrict international movements of factors and goods which enables governments to engage in costly redistribution policies.

Migué's fundamental insight is a simple but powerful one: "free trade equals federalism." With free trade, governments are thrown into competition with one another for the patronage of consumers who can move their resources (tax bases) to those political jurisdictions that provide the best value for their tax dollars. The advantage that organized interests have over the general taxpaying public at communicating through the political process becomes less important as exit (as opposed to voice) options become more important.

Some have argued that attempts to restrict government power are likely to be futile at best and inefficient at worst. If, for example, a spending limit is imposed on government, it will employ some other means of controlling resources (e.g., regulation). And this alternative will probably be a less efficient way of directing resources than taxing and spending, with the result being worse than before. That argument is surely overstated for reasons Marilyn Flowers (1992) has put forth. But it is true that most attempts to control government with one policy initiative (e.g., spending limits) weaken the ability to control government

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

with another (e.g., regulatory reform). A policy of free trade, as Migué points out, imposes restraint on government while strengthening, rather than weakening, other means of restraining government power. In his words:

> Without the power to close their borders to foreign competition, national governments discover that most other tools formally available to them to pursue political objectives are of little value, because resort to such instruments becomes too costly.

Migué develops his proposition concerning the close relationship between federalism, free trade, and control over government excesses with theoretical discussion and illustrative examples. He does a nice job, for example, challenging the conventional view that tax efficiency requires closing off opportunities to escape taxes. Following Geoffrey Brennan and James Buchanan (1980), Migué points out that this view assumes that government requires a given amount of revenue, and the objective is to raise that amount of revenue with the least distortion to the economy. The problem is that public demands are as insatiable as private demands, and closing off opportunities to escape taxes creates opportunities for government growth that are not overlooked. Free trade and federalism offer escape opportunities that impose discipline on political appetites.

Migué supports his theoretical discussion with examples of the type of taxes one would expect to become more prominent under a regime of free trade and federalism (taxes that do a better job pricing the use of government services), and of the responses one would expect from government to counter the discipline of free trade and federalism (such as policy harmonization in the European Union).

Though the idea is not explicitly developed, Migué's work hints at the possibility of a virtuous circle of freer trade and more effective constraints on government power, a circle that may already have been initiated by technological advances. Despite the best efforts of politicians, public bureaucrats, and their special-interest clients, technological advances—and the economic globalization these advances are helping to create—are making it easier for productive citizens to escape the clutches of those who specialize in transfer activity. Just as businesses are increasingly facing global competition from other businesses, governments are finding themselves increasingly having to compete with other governments, and finding that the most effective way of meeting this competition is with policies that facilitate the production of wealth instead of with policies that redistribute wealth (see McKenzie and Lee 1991). The liberalization of international trade, while admittedly advancing with less haste and more contaminating side agreements than is ideal, is an indication of the competitive discipline being imposed on governments. And with freer trade comes a more effective federalism, which, as Migué argues, makes governments more responsive to the general interest served by efficiency and less responsive to the special interests served by redistribution. This

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

can lead to even more pressure to liberalize trade, and the virtuous circle continues.

The point is not that economic nirvana is right around the corner. But the argument that positive pressures are at work in the political economy is worth making, both because it is true and because it takes some of the pessimism out of public choice analysis. Alfred Marshall once said, "Without hope there is no enterprise," and few enterprises are more important than controlling government. Migué has contributed admirably to that enterprise.

Although *Federalism and Free Trade* is a small book, it offers the reader a big return in insight and optimism.

Dwight R. Lee
University of Georgia

References

Brennan, G.G., and Buchanan, J.M. (1980) *The Power to Tax: Analytical Foundations of a Fiscal Constitution.* Cambridge: Cambridge University Press.

Flowers, M. (1992) "The Political Economy of Mandated Spending." *Cato Journal* 12(2): 337–47.

McKenzie, R.B., and Lee, D.R. (1991) *Quicksilver Capital.* New York: Free Press.

Why Perestroika Failed: The Politics and Economics of Socialist Transformation
Peter J. Boettke
London and New York: Routledge, 1993, 199 pp.

This compact, thoughtful, and ideologically charged book might as easily have been entitled *Why Socialism Fails*. It is an elaborate argument about economic theory, not an account of the day-to-day reality of Mikhail S. Gorbachev's reforms. It was written in the placid atmosphere of the New York University department of economics, not the grittier settings where Gorbachev's stop-and-go programs worked their tumult. The book's index notes 11 pages citing F. A. Hayek, more than the combined citations of the principal Soviet architects of economic reform: Leonid Abalkin, Abel Aganbegyan, Stanislav Shatalin, and Grigory Yavlinsky. Nikolai Ryzhkov, prime minister for most of the *perestroika* period, merits just one mention; his successor, Valentin Pavlov, does not appear. For 146 pages of text there are 47 pages of notes, but Russian sources are cited only in translation. The prose will gladden the heart only of a fellow economist: "In a large number setting individuals treat rules as parametric, similar to how agents within the perfectly competitive model of general equilibrium treat price as a given."

Yet to a former Moscow correspondent whose observation of perestroika was up-close and innocent of economic theory, Boettke's book seems a persuasive explanatory frame for the Gorbachev era. Boettke is

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

absolutely right that Gorbachev's attempts at "renewing socialism" were doomed not by tactical errors but by fundamental contradictions—that favorite Marxist notion—in the very design of the Soviet socialism to which the Soviet leader clung.

Long after Boris Yeltsin had swept his meaty arm at the shelves of a Houston supermarket, declaring, "Even the Politburo doesn't have such choice," Gorbachev was still repeating his mantra: "the people's socialist choice in 1917." It was the profound irony of the era that Gorbachev remained publicly oblivious to the explosion of historical publications that he had permitted and that amply demonstrated, among many other things, that the Great October Socialist Revolution was by no means a "choice" exercised by "the people."

Boettke correctly places the economic portion of Gorbachev's reforms in the tradition of periodic Soviet attempts to rev up the economy without abandoning central planning: the New Economic Policy of Lenin, beginning a scant four years after the Bolshevik coup; Khrushchev's *sovnarkhoz* reforms; and the Brezhnev-era reforms of 1965, 1973, and 1979. Markets, Boettke argues, constitute not theory but reality, like gravitation or photosynthesis. "The peculiar art of Soviet economic policy was to balance an ideological hatred of market relations, which justified the Party's privileged position in society, with the reality of allowing enough market production and exchange so that the Party's ability to extract rents was not threatened," Boettke writes. "Soviet leaders were chosen for their ability to uphold the fiction that the fictional reality of communism was not fictitious." It is an eloquent phrasing of the matter, and one that will ring true to anyone who attended the press briefings that followed Central Committee plenary meetings in the late 1980s. All present pretended that the crucial questions were about the details of the reform plan of the moment; yet it was evident that the only way out of the quagmire led through market prices and private property. The corpse of communist economic theory had already begun to decompose, but at those press briefings no one was so impolite as to remark upon the stench.

For all the elegance of his theory, however, Boettke's book seems to me misleading in certain respects. He overstates Soviet economic failure. The true collapse of the economy was caused by, and did not precede, the halting steps of perestroika. The first hints of price hikes, in 1990, resulted in hoarding that caused the perennial Soviet weaknesses of long lines and empty shops to become intolerable. Industrial breakdown became acute as a result of strikes, ethnic conflict and—most important—the sudden freedom of managers to pick and choose among suppliers and customers. Boettke argues that you cannot leap the chasm between Soviet socialism and a market economy in a series of smaller jumps. But that is precisely what Gorbachev attempted.

Boettke likens the Soviet economy to "a well-developed Third World economy." But that familiar comparison blurs interesting differences. The Moscow Metro, Aeroflot, Soviet Olympic teams, and music conservatories on the one hand, and the Soviet military and KGB on the other—

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

these were not Third World institutions. They were testimony to what a planned economy could achieve by the uneconomic application of massive resources. Only the ideologically blind could compare public health in, say, Uzbekistan and Afghanistan and fail to see the Soviet achievement. Western and Soviet critics in the late 1980s decried the infant mortality rate; yet the Soviet rate was approximately where the U.S. rate had been in 1960.

Indeed, Boettke's persuasive argument for the doomed nature of the Soviet system raises a question: if it was doomed, why did it collapse in 1991 and not in 1931, 1951, or 1971? His book ignores the dynamic of the world economy that forced the issue: the information revolution. Central planning and police powers were suited to the construction and operation of industrial mega-plants such as the Magnitogorsk steel combine. They were utterly incompatible with the more subtle requirements of a microchip economy whose skilled workers must freely communicate via computer, modem, and fax machine. In the late 1980s the Soviet Union produced twice as much steel as the United States—but the United States had more than 100 times as many personal computers in use. Had Chernenko been a little more spry, certainly the U.S.S.R. could have dragged itself on for a few more years, perhaps even to the turn of the century. But it was the postindustrial turn in the global economy that prompted Gorbachev's fatal experiment.

By focusing exclusively on Gorbachev's economic fiddling, Boettke slights what was unquestionably Gorbachev's greatest achievement: *glasnost*. As Boettke notes, sneaking timid elements of the market into the Soviet economic scheme was hardly novel. But freeing the media to the degree that Gorbachev did was a bold break not only with Soviet practice of seven decades but even of the older Russian tradition. (Boettke's assessment that "Gorbachev liberalized government restrictions in some ways" is stingy.) Highly educated, curious Soviet people took the freedom up with such a vengeance that glasnost—far more than economic perestroika—proved fatal to the regime. Ironically, perhaps Gorbachev's very temporizing, his clinging to Soviet rhetoric even as Soviet power slipped away, ensured the end of the old regime. A forthright anti-communist, à la Yeltsin after 1989, would certainly have been overthrown before he could have made an impact. A conservative would never have dared ease controls on the press and history. Gorbachev's tendency to dress the repudiation of Leninism in the garb of renewed Leninism allowed the overseers of the army and KGB to slumber until it was too late.

The stunning reversal of the Russian economic problem—from empty stores and long lines to overflowing stores and no lines—has certainly proven Boettke's thesis that structural economic factors, and in particular fixed prices, were at the core of Soviet economic failure. Westerners who argued that Russians had some innate allergy to markets and democracy were mistaken. "If it is understood that the ugliness that occurred was due to institutional failings, then institutions can be established to guard against its ever happening again," Boettke writes.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

True enough. But institutional change takes time, as the current state of Russian capitalism attest. Boettke is too casual in his dismissal of the significance of inflation and ethnic conflict. And his call for "a new, but workable utopia," may not readily be embraced by the Russian people. Utopia is a word that scares Russians these days. They have had enough of utopia. They want what they call, prosaically and without reference to either Marx or Hayek, "a normal country."

Scott Shane
Baltimore Sun

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

JOURNAL OF MONEY, CREDIT, AND BANKING

Widely read and cited by researchers and policymakers, the *Journal of Money, Credit, and Banking* is a primary economics journal reporting major findings in the study of monetary and fiscal policy, credit markets, money and banking, portfolio management, and related subjects. Editors: Stephen G. Cecchetti and Paul D. Evans, Ohio State University.

Recent and Forthcoming Articles...

The Foreign Exchange Risk Premium: Is It Real? *Craig S. Hakkio and Anne Sibert*

The Relationship between Capital and Earnings in Banking *Allen N. Berger*

Near-Rationality, Heterogeneity, and Aggregate Consumption *Richardo J. Caballero*

Fighting Inflation with High Interest Rates: The Small Open Economy Case under Flexible Prices *Guillermo A. Calvo and Carlos A. Végh*

Core Deposits and Physical Capital: A Reexamination of Bank Scale Economies and Efficiency with Quasi-Fixed Inputs *William C. Hunter and Stephen G. Timme*

Liquidity Constraints and Intertemporal Consumer Optimization:Theory and Evidence from Durable Goods *Eun Young Chah, Valerie A. Ramey, and Ross M. Starr*

PLUS...

Liquidity, Monetary Policy and Financial Remediation—papers and comments from a conference sponsored by the Federal Reserve Bank of Cleveland. A *JMCB* Special issue supplement (November 1995)—free with order for subscription.

Published quarterly by the **Ohio State University Press**
1070 Carmack Road, Columbus, Ohio 43210-1002 614/292-6930
One-year subscription: $40 Individual $100 Library $25 student

---

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

JOURNAL OF LABOR RESEARCH

Symposia and Selected Articles Forthcoming in Volume XVII, 1996

The Dunlop Commission Report
Herbert R. Northrup, John S. Irving, Barry T. Hirsch, Jeffrey C. McGuiness, John T. Delaney, Harold Coxson, Edward B. Miller, John T. Addison, Dwight R. Lee, Thomas J. DiLorenzo, William J. Dennis, Jr., Morgan O. Reynolds, Bruce E. Kaufman

Comprehensive Corporate Campaigns
Peter Hoekstra, Charles R. Perry, Herbert R. Northrup, Thomas J. DiLorenzo, Joseph Uehlein, C.R. Gangemi, Richard Vigilante, George Hildebrand, Jeffrey C. McGuiness

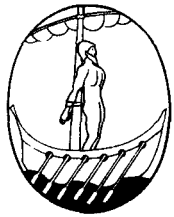
The Long-Run and Short-Run Impacts of Global Competition on U.S. Union Wages	Stephen B. Blumenfeld and Mark D.Partridge
Unemployment, Union Density, and Wages	Terry Ashley and Ethel B. Jones
Dispute Resolution Mechanisms and Teacher Bargaining Outcomes	Michael A. Zigarelli
Financial and Political Resources of Nine Major Public Sector Unions in the 1980s	Marick F. Masters and Robert S. Atkin
Does Strike Insurance Matter? Evidence From The Airline Industry's Mutual Aid Pact	Pierre-Yves Crémieux
Information, Disinformation, and Union Success in Certification and Decertification Elections	Victor G. Devinatz and Daniel P. Rich
Bonuses, Unions, and Labor Productivity in South Korea	Michael Byungnam Lee and Yinsog Rhee
The Effect of the United Mine Workers of America on the Probability of Severe Injury in Underground Coal Mines	Jack Reardon
Compliance with ERISA's Disclosure Provisions	Mitchell Langbert
Labor Laws: Then and Now	Lowell Gallaway and Richard Vedder
Explaining the Unionism Gap: Public-Private Sector Differences in Preferences for Unionization	Jack Fiorito, Lee P. Stepina, and Dennis P. Bozeman
Employment Growth in the Temporary Help Supply Industry	Karylee Laird and Nicolas Williams
Employee Involvement Programs: The Noninvolvement of Unions Revisited	Robert E. Allen and Kathleen L. Van Norman
A Model of Monopoly and "Efficient" Unions with Endogenous Union Coverage: Positive and Normative Implications	Brian Chezum and John Garen
External Legitimacy of Unions: Trends in New Zealand	Raymond Harbridge and Anthony Honeybone

The **Journal of Labor Research** is published quarterly by the Department of Economics of George Mason University. Subscriptions are on a calendar year basis and subscribers will receive all issues for the current year. Annual subscription rates are: individuals, $35.00; academic libraries, institutions, government agencies, and business firms, $110.00. Foreign subscribers add $15.00 to cover additional postage. All remittances should be made payable to **Journal of Labor Research,** Department of Economics, George Mason University, Fairfax, Virginia 22030.

Editorial correspondence and manuscripts should be sent to James T. Bennett, Editor, Journal of Labor Research, Department of Economics, George Mason University, Fairfax, VA 22030. Manuscripts should be submitted in triplicate with a 100-word abstract. A submission fee of $20.00 for nonsubscribers must accompany all manuscripts. Papers accepted for publication must conform to JLR style requirements; style sheet available on request.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

CONSTITUTIONAL POLITICAL ECONOMY



Advisory Editor
JAMES M. BUCHANAN
Nobel Laureate

Editors
VIKTOR J. VANBERG
RICHARD E. WAGNER

"Yet, although the problem of an appropriate social order is today studied from the different angles of economics, jurisprudence, political science, sociology, and ethics, the problem is one which can be approached successfully only as a whole."

Friedrich A. von Hayek

The journal CONSTITUTIONAL POLITICAL ECONOMY is a forum for papers in the broad area of constitutional analysis, which lies at the intersection of several approaches in modern economics. All of these approaches share an interest in systematically integrating the institutional dimension—the study of political, legal, and moral institutions—into economic analysis.

Though its primary locus is in economics, CONSTITUTIONAL POLITICAL ECONOMY has an explicit interdisciplinary orientation. One of its aims is to invite and encourage interdisciplinary exchange, including contributions from the various social sciences, philosophy, law, etc. Accordingly, CONSTITUTIONAL POLITICAL ECONOMY places particular emphasis on articles which, in style and substance, are suitable to this end. Theoretical papers and empirical studies are published, as well as contributions to constitutional policy issues. Reviews of pertinent books are also included.

To receive more information about CPE or to submit a paper, write to the editorial office:

Constitutional Political Economy
Center for Study of Public Choice
George Mason University
Fairfax, Virginia 22030-4444

To subscribe to CPE please enclose your check or money order with your subscription.

Please send a sample copy ☐
Please enter my individual subscription $27 ☐
Please enter my institutional subscription $65 ☐
(Subscribers outside the U.S. add $9 for air delivery or $6 for surface mail.)

Name _____

Institution _____

Street _____

City/State/Zip _____

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

Cato Institute

Founded in 1977, the Cato Institute is a public policy research foundation dedicated to broadening the parameters of policy debate to allow consideration of more options that are consistent with the traditional American principles of limited government, individual liberty, and peace. To that end, the Institute strives to achieve greater involvement of the intelligent, concerned lay public in questions of policy and the proper role of government.

The Institute is named for *Cato's Letters*, libertarian pamphlets that were widely read in the American Colonies in the early 18th century and played a major role in laying the philosophical foundation for the American Revolution.

Despite the achievement of the nation's Founders, today virtually no aspect of life is free from government encroachment. A pervasive intolerance for individual rights is shown by government's arbitrary intrusions into private economic transactions and its disregard for civil liberties.

To counter that trend, the Cato Institute undertakes an extensive publications program that addresses the complete spectrum of policy issues. Books, monographs, and shorter studies are commissioned to examine the federal budget, Social Security, regulation, military spending, international trade, and myriad other issues. Major policy conferences are held throughout the year, from which papers are published thrice yearly in the *Cato Journal.* The Institute also publishes the quarterly magazine *Regulation.*

In order to maintain its independence, the Cato Institute accepts no government funding. Contributions are received from foundations, corporations, and individuals, and other revenue is generated from the sale of publications. The Institute is a nonprofit, tax-exempt, educational foundation under Section 501(c)3 of the Internal Revenue Code.

CATO INSTITUTE
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED

$8.00

VIRTUE AND THE LAW

Law's stock in trade is the use of collective might, and the sanctions that it imposes must be reserved for the most serious of social violations—which typically involve using force and not keeping serious contractual engagements. A legal system is not a complete social system, and we should not reflexively invoke legal remedies to enforce whatever conduct we think to be socially desirable. The uncritical reliance on coercion drives out informal sanctions and curtails possibilities for displaying the personal virtues we all prize. The most that we can expect from a legal system is to create a domain that allows the virtuous a free sphere of action to do good deeds.

—Richard A. Epstein

Simple Rules for a Complex World
(Cambridge: Harvard University Press, 1995), 326.

LICENSED TO UNZ.ORG
ELECTRONIC REPRODUCTION PROHIBITED